|  |
| --- |
|  |
| CS440 Project Report |
| Second Submission |
|  |
| **Naveen Kumar Lekkalapudi** |
| **5/3/2013** |

|  |
| --- |
|  |

# Content Layout

The submission consists of 3 files and a directory:

1. CS440 Project Report
2. CS440 Project Analysis
3. Script.pig
4. 10yrdata

CS440 Project Report is the current word file.

Script.pig is the source file containing the pig script which used to obtain the data.

CS440 Project Analysis is an excel file which contains the spread sheets of data. It has total of 25 spread sheets listed as follows:

1. Temp ptable contains Average Temperature analysis
2. Temp 10yr contains Analysis of 2000-2009 temperature data
3. Dewp ptable contains Average Dew Point analysis
4. Slp ptable contains Average Sea Level Pressure analysis
5. Stp ptable contains Average Station Level Pressure analysis
6. Visibility ptable contains Average Visibility analysis
7. Wind Speed ptable contains Average Wind Speed analysis
8. Mxspd ptable contains Max speed analysis
9. Gust ptable contains Max gust in wind analysis
10. Max ptable contains Maximum temperature analysis
11. Min ptable contains Minimum temperature analysis
12. Prcp ptable contains Average Precipitation analysis
13. Sndp ptable contains Average Snow Depth Analysis

The next 12 tables contain raw output of data obtained from the script after running of Hadoop of amazon.

10yrdata has 4 files and 2 directories as follows:

1. Bshscriptforpig.sh 🡪used to run the Script.pig across the past 10 years
2. Script.pig
3. BshscriptforR.sh 🡪used to run Rscript.R across the output generated from the Bshscriptforpig.sh
4. Rscript.R
5. Rawoutput10yr contains raw output generated from pig
6. Graphoutput10yr contains 6000 graphs (approx.) for years 2000-2009 for each state for each month, the average temperature.

# Summary

I have learnt new things in my experience with hadoop. I have used pig programming to interact with the big dataset. It seemed to be very effective and yet simple to work with. I have also tried using python and java but comparatively pig seemed a bit easier.

Amazon Elastic Map Reduce has a wide range of tutorials listing almost every way to interact with it in form of tutorials and videos. They have always been handy when I was stuck at any problem.

Below are a list of problems I have faced and how I was able to solve them:

### Connecting to hadoop

I initially faced inconsistencies while connecting to hadoop, as it listed my keys are public or permission is denied even when the credentials were correct. Just a new security key solved my problem.

### Pig Upgrade, may be

The pig interface in tutorials is different from the pig interface that is currently provided, it may be because of the upgrade from an older version of pig but the newer versions tend to give more “INFO” and “WARN” messages which was quite annoying.

### Pig programming, regular expressions

I have used regular expressions to extract the raw data into tuples which was an efficient but very time consuming. I had to try a variety of combinations to make it work.

Once I figured out how to use amazon services, it was an easy walk to get the script done and shoot the output into the s3 bucket.

I used Microsoft excel to perform the analysis, and generate the graphs.

Note: All units of data collected are as per the national climatic data center code.

# Procedure

I have used elastic map reduce service of amazon which gives an option to execute a pig script over a set of virtual nodes.

Input: s3://cs440-climate/gsod/

Script Location: s3://cs440-nalekkalapudi/script.pig

Output: s3://cs440-nalekkalapudi/output

# Analysis

The analysis can be summarized as following:

1. Temperature analysis from temp ptable
   1. The average temperature of United States over the years 1931-2009 is 53.97 approx.
   2. Hawai seems to be the hottest state with an average temperature of about 74.45.
   3. Alaska seems to be the coolest state with an average temperature of about 31.01.
2. Temperature analysis from past 10 years
   1. Few states seems to record a consistent increase in temperature such as FL, GA, etc.
   2. Few states seems to record a downward slope ME, NE, etc.
   3. These rapid changes might be the effect of global warming and imbalance of atmospheric gases.
3. Dew Point Analysis
   1. Minimum Average Dew Point recorded is in Montana on January of 1942 with a value of about -18.6.
   2. There is a fair amount of consistency in the dew as most of the states have consistent dew points.
4. Sea Level Pressure
   1. Interestingly maximum Sea Level Pressure is also recorded in Montana on January of 1942 of about 1036.1.
   2. The average sea level pressure is around 1016.508 over the years 1933-2009.
5. Station Level Pressure
   1. The average station level pressure is 972.41 over the years 1938-2009.
6. Visibility
   1. Visibility has been rapidly decreasing; it may be because of extreme difference in climatic conditions due to the raise of industries.
   2. On an average 15.7 miles can be seen.
   3. The visibility used to be around 15 miles in mid 1950s and 1960s which has fallen to around 9 miles in late 2000.
7. Average Wind Speed
   1. An average wind speed of 7.47 is recorded throughout the years 1931-2009.
8. Maximum Wind Speed
   1. A maximum speed of 96.9 is recorded at several places across united states over the period of time.
9. Gust
   1. A maximum wind gust of 135 is recorded in NH in October of 2006.
10. Maximum Temperature
    1. A maximum temperature of 131 degree Fahrenheit is recorded in California in September of 1992.
11. Minimum Temperature
    1. A minimum temperature of -95.8 is recorded degree Fahrenheit is recorded in Alaska in March of 1982.
12. Precipitation
    1. A highest precipitation of 90.23 is recorded in Michigan in January of 1957.
13. Snow Depth
    1. A maximum average snow depth of 109 inches is recorded in Washington in April of 2002.

These are a few of the shocking facts extracted by the analysis of weather data from 1929 to 2009.

From the graphs obtained from the 10 years, we can notice that a common pattern of sinusoidal wave can be seen over the months march to august in most states. The temperature rises for few days and then falls again then rises.

On an average, we could see a slight increase in temperature of about 1-2 degrees over the past few decades and also severe inconsistencies which are rendering climate in to a state of instability. These might the side effects of increasing pollution and depleting natural resources.