RESULTS AND DISCUSSIONS

Q1) Which state has the most monitoring sites across the United States?
Note: a site is identified by the combination of the state code, county code and site number

A - **The answer I got after executing this job is that California has the maximum number of sites with 179 sites**.
I reached this answer by storing the unique sites(stateCode + countyCode + siteNum) and keeping track of them so that I am not sending redundant data over the network. I emit from my mapper, <stateName, siteID>, once per unique site per map task. On the reducer end I again filter out duplicate sites against the keyed state name whilst updating the maximum site number and the state containing them. Since I have one reducer instance and I emit the final answer from the cleanup method.

Q2) Does the East Coast or West Coast have higher mean levels of SO2?
Note: there are a total of 4 and 16 states in the West Coast and East Coast, respectfully

A - The results I output from this job was as follows:
The average SO2 levels of the West Coast states are 2.235199107768116 ppb and the average of the East Coast states are 6.117746126946715 ppb!
Therefore, we can see that the SO2 levels were higher along the East Coast. For this task, I filter by state names and emit from the mapper <statename, sum:count>; which is a key of state name(along with a Eastern/Western qualifier appended to it) and value of the sum of SO2 observations and their counts as a colon delimited string. On the reducer end I add up all the sums and the counts and calculate the mean of the west coast and east coast states.

Q3) What time of day (GMT) has the highest SO2 levels between 2000 – 2019? Capture the mean SO2 levels for each hour (GMT) over all 20 years to justify your answer.

A - The results I output from this job was as follows:
The highest average SO2 for a given hour of day between 2000 - 2019 is for hour number 16 of the day and the mean SO2 level for that hour is 3.569352595827633
I have also stored the hourly means for every hour of day in my hdfs. For this task, I emit from my mapper a key of the hour of day and the value of the sum of SO2 observations and their counts as a colon delimited string for each hour. On the reducer end I add up all the sums and the counts and calculate the mean of the SO2 levels for every given hour of day.

Q4) Has there been a change in SO2 levels over the last 40 years? Capture the mean SO2 levels for each year to justify your answer.

A - Some sample results from this job are as follows:
The average SO2 for the year 1980 is 9.20398877716909
The average SO2 for the year 2019 is 0.8593651379244132
I have stored the results for all years in my hdfs as well. Looking at that I can make a substantial claim that there has been a steady decrease in SO2 levels over the last 40 years. There has been a decrease of over 80% in the SO2 levels from 1980 to 2019. For this task, I emit from my mapper a key which is the year of observation and the value of the sum of SO2 observations and their counts as a colon delimited string for each year. On the reducer end I add up all the sums and the counts and calculate the mean of the SO2 levels for every year.

Q5) What are the top 10 hottest states for the summer months (June, July, August)? Capture the mean temperature levels for the summer months (GMT) to justify your answer.

A - Some sample results from this job are as follows:
Hottest State - The average temp for the summer month in state Arizona is 85.23638377778865F
10th Hottest State - The average temp for the summer month in state Oklahoma is 80.16688561268104F
I have stored the summer month temperatures for the past 20 years for all states in my hdfs. The top 10 hottest states in the summer months are Arizona, Puerto Rico, Arkansas, Mississippi, Virgin Islands, Oklahoma, Texas, Nevada, Florida and Louisiana(not in that order). For this task, I emit from my mapper the state name as the key and the value of the sum of temperature observations and their counts as a colon delimited string for each state, while considering a record if and only if the months is in the range 6-8. On the reducer side I average the temperature observations and order them by utilizing a wrapper over the state name and the temperature and using Collections.sort in the cleanup method to write the top 10 states to file.

Q6) What are the mean SO2 levels for the hottest states found in Question 5?

A - Again, showing sample results,
The yearly SO2 mean (all 40 years) for the state georgia is 4.008332259583279
The yearly SO2 mean (all 40 years) for the state nevada is 0.7840678884566422

I store the hottest state found from previous job and filter SO2 only on those states and emit from the mapper (state name + year) as the key and the sum and count of state + year observations which I tally in the reducer thus having access to yearly averages and well as all years averages which I finally output. The full output(for all 10 states) is safely stored in hdfs.