3.1 Supporting Documentation

Team 14

Cameron Brightwell, Kevin Conner, and Mridul Banik

Question 1:

First we read the merged csv file. From the csv file we the get AQI score. We needed to convert epochtime to day. Then we pass day, aqi scores as key, value pair from mapper to reducer. In reducer, we took summation of all AQI score of corresponding day. Then we sort the file and output that into HDFS.

Best Day : Monday

Worst Day : Friday

Question 2:

First we read the merged csv file. From the csv file get AQI score. We needed to convert epochtime to Month. Then we pass month, AQI scores as key, value pair from mapper to reducer. In reducer, we took summation of all AQI score of corresponding months. Then we sort the file and output that into HDFS.

Best Month: February

Worst Month : May

Question 3:

First we read the merged csv file. We needed to convert epochtime to year and took aqi score only from the year 2020. We kept the information of County and it's corresponding state in key and aqi score in value and passed that to reducer. In reducer, we took average of all aqi score of corresponding county. Then we kept only the Ten best counties and output that into hdfs. Here's the output of the 10 best county based on average AQI score.

64 Kings County, California

61 Fresno County, California

56 San Bernardino County, California

54 Tulare County, California

53 Butte County, California

51 Sutter County, California

50 Sacramento County, California

49 San Joaquin County, California

47 Tehama County, California

45 Placer County, California

Question 4:

Mapper maps all aqi date by county and state. The reducer calculates the average for each county in the data set for 2020. A TreeMap is used at setup and cleanup to sort the counties based on their averages by using the average as the key. When the TreeMap size is greater than 10, possible ties in avg AQI are considered before being removed. Here's the output of the 10 worst counties based on average AQI score.

12 Fremont County, Colorado

14 San Miguel County, Colorado

15 Routt County, Colorado

16 Putnam County, Florida

17 Skagit County, Washington

18 Rockdale County, Georgia

19 Thurston County, Washington

20 Sumter County, Georgia

21 Whatcom County, Washington

22 Wakulla County, Florida

Question 5:

Used the java.time library to get the week numbers from 1-52 or sometimes 1-53 for the WEEK_OF_WEEK_BASED_YEAR format. All AQI data mapped to a county and state as the key. This will consider counties with the same name in different states as separate counties. The reducer used a TreeMap at cleanup and setup to sort the counties based on their best one week change in AQI score.

Question 6:

Two mappers are used to map two different csv input files. One gets total AQI by state. A ranking score from 1 to total number of states in the data is assigned based on highest to lowest AQI for each state. Similarly, the other mapper assigns a rank to states based on number of in-service oil refineries. The reducer combines the rank to form a combined ranking for all states in the data set. Higher ranks are given for higher AQI scores and greater number of in-service refineries and vice versa.