# Homework 3

ANALYZING AIR QUALITY DATA COLLECTED ACROSS THE UNITED STATES USING MAPREDUCE
VERSION 1.0

DUE DATE: Wednesday April 20th, 2022 @ 5:00 pm

## OBJECTIVE

The objective of this assignment is to gain experience in developing MapReduce programs. As part of this assignment, you will be working with data collected from the EPA's Air Quality System (AQS). You will be developing MapReduce programs that parse and process recordings of temperature and criteria gas levels at various outdoor monitors.

You will be using Apache Hadoop (version 3.1.2) to implement this assignment. Instructions for accessing datasets and setting up Hadoop clusters will be released in Canvas alongside this assignment.

This assignment accounts for **20% of your course grade** and may be modified to clarify any questions (and the version number incremented), but the crux of the assignment and the distribution of points will not change.

## 1 Cluster setup

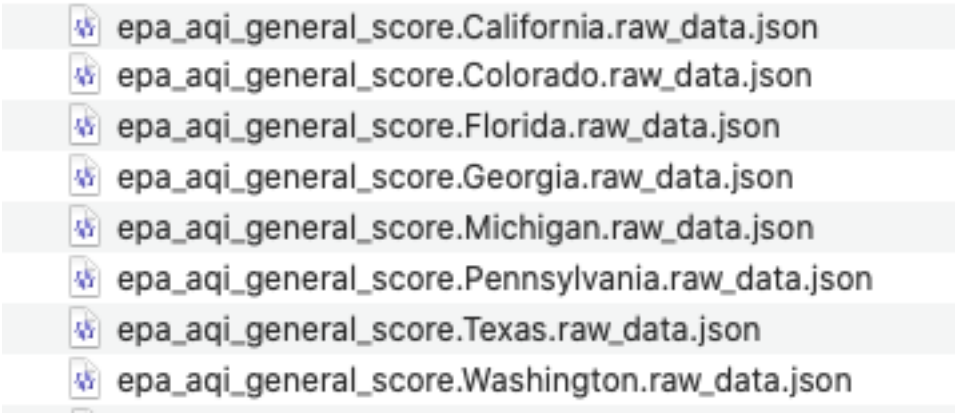
As part of this assignment your team is responsible for setting up your own Hadoop cluster with HDFS running on every node. You are also required to stage your own datasets. Your programs will process the staged datasets; data locality will be preserved by the MapReduce runtime.

## 2 Air Quality Dataset

The dataset contains daily measurements of air quality (AQI) readings from various monitors around the United States. Records are stored in separate CSV files. The file name consists of the type of data and the state. For example, records for Colorado are stored in a file named `epa_aqi_general_score.Colorado.raw_data.json`. The readings are stored in JSON format containing: a unique id, epoch time, gisjoin id and the AQI score. The total number of records per files depends on the number of counties in each state.

The assignment will use the AQI data from the following states: California, Colorado, Florida, Georgia, Michigan, Pennsylvania, Texas, and Washington. The datasets can be downloaded from the Sustain website (https://urban-sustain.org) using the Data Download service, the name of the dataset is EPA AQI Score.

The below screenshot will display the files you should use, downloaded from the Urban Sustain Website, for this assignment.




The data must be processed into a single file, in which each line is a unique entry. Students may choose to process the data themselves or use the script provided by us. At the end you will be left with a single file with 2,120,911 lines.

## 3 Analysis of Air Quality Data

You should develop MapReduce programs that process the AQS dataset to answer the following questions. We need to a

| | |
|---|---|
| **Q1.** | What were the best and worst Days of Week for AQI scores? |
| **Q2.** | What were the best and worst Months for AQI scores? |
| **Q3.** | Which 10 counties had the best average AQI scores for the year 2020? |
| **Q4.** | Which 10 counties had the worst average AQI scores for the year 2020? |
| **Q5.** | List the biggest one week change for each county in the dataset. |
| **Q6.** | Do your own analysis. For this question, you are no longer restricted to using AQI data from 8 states – you can, if you'd like, use data from all 50 states.<br>**Note:** This will be graded on complexity, you may use additional datasets to supplement this analysis. |

### 3.1 Supporting Documentation

You should include a PDF report that substantiates the results from your analysis. This document should specify (a) the answers from the aforementioned questions as well as (b) a description elaborating on the methods used to get to those answers. Your descriptions should only be a few sentences per question (no more than 4 sentences each is required). If unable to reach an answer to a specific question, then include a description for how you would have approached the problem. Please only include a PDF document and not Word, OpenOffice, or Google Docs please.

## 4 Additional Requirements

Grading will be conducted by interview, and it is important that you are able to explain the method you used to get your answer and why you believe that method accurately answers the question asked.

Try to design your MapReduce jobs as elegantly as possible. This means minimizing the number of jobs and the amount of data transferred between each job. Minimizing the amount of data transferred between the mapper and reducer within each job is also important as it significantly impacts the amount of time the job will take to run.

## 5 Additional Requirements

There will be an **18-point deduction** if any of the restrictions below are violated.

1. You should not implement this assignment as a stand-alone program.
2. You should not implement this assignment using anything other than Hadoop MapReduce. Implementing your own framework or using a 3rd party framework (that is not Hadoop) to implement this assignment is not allowed.

## 6 Grading

Homework 3 accounts for 20 points towards your final course grade. The point distribution for this assignment is listed below.

**Point Breakdown:**

| | |
|---|---|
| 3 points: | Correctly configured Hadoop cluster |
| 10 points: | 2 points each for the correct answer for questions Q1-Q5 |
| 4 points: | Q6 |
| 3 points | Substantiate results for Q1-Q6 in a report (please save your document as PDF) |

## 7 Milestones:

You have 4 weeks to complete this assignment. The weekly milestones below correspond to what you should be able to complete at the end of every week.

Milestone 1: You should have your HDFS/MapReduce cluster configured. You should be able to read data from your HDFS cluster into a MapReduce program and write data from a MapReduce program back to the cluster. You should also be able to read the AQS dataset from the shared HDFS cluster.

Milestone 2: Develop MapReduce programs to answer Q1-Q4 and write answers your answers to your local HDFS cluster.

Milestone 3: Complete the MapReduce implementations for Q5 & Q6. Put the final touches on your report to substantiate your results from Q1-Q6.

## 8 What to Submit

Use **CANVAS** to submit a single .tar file that contains:
- all the Java files related to the assignment (please document your code)
- the `build.gradle` file you use to build your assignment
- a README.txt file containing a manifest of your files and any information you feel the TAs needs to grade your program.

E-mailing the codes to the Professor, GTA, or the class accounts will result in an automatic 1 point deduction.

**Filename Convention:** All classes should reside in a package called **`cs455.aqi`**. The archive file should be named as TeamName-HW3.tar. For example, if your assigned team name is Alpha then the tar file should be named Alpha-HW3.tar.

## 9 Version Change History

This section will reflect the change history for the assignment. It will list the version number, the date it was released, and the changes that were made to the preceding version. Changes to the first public release are made to clarify the assignment; the spirit or the crux of the assignment will not change.

| Version | Date | Comments |
|---|---|---|
| 1.0 | 3/22/2022 | First public release of the assignment. |