Homework 3: Programming Component

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

DUE DATE: Wednesday April 15th, 2020 @ 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 hourly recordings of temperature and criteria gas levels at various outdoor monitors between 1980 and 2019.

You will be using Apache Hadoop (version 3.1.2) to implement this assignment. Instructions for accessing datasets and setting up Hadoop clusters are available on the course website.

This assignment 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 you are responsible for setting up your own Hadoop cluster with HDFS running on every node. We will be staging datasets on a *read-only* cluster. You should use your ***own*** cluster to write outputs produced by your MapReduce programs. MapReduce clients will be able to access namespaces of both clusters through Hadoop ViewFS federation. Your programs will process the staged datasets; data locality will be preserved by the MapReduce runtime.

2 Air Quality Dataset

The dataset contains hourly measurements of meteorological temperatures and sulfur dioxide (SO2) 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 year. For example, records for 2001 are stored in a file named `hourly_42401_2001.csv` and `hourly_TEMP_2001.csv` for SO2 criteria gases and meteorological data, respectfully. Each line in a file corresponds to a single record comprised of comma separated fields. There are approximately 360 million records in the entire dataset that total 65 GB. The dataset is available under the directories `/data/gases` and `/data/meteorological` within the shared HDFS.

The complete documentation including the data dictionary for the dataset is available online at https://www.epa.gov/outdoor-air-quality-data. The table in the following section summarizes the fields. Fields in the following table are appearing in the same order as in a record.

2.1 Meteorological and Criteria Gas Data Fields

Index	Field Name	Description
1	State Code	FIPS code of the state
2	County Code	FIPS code of the county within a state
3	Site Num	Unique number within the county identifying the site
4	Parameter Code	AQS code corresponding to the parameter measured
5	POC	Parameter Occurrence Code for different instruments
6	Latitude	Angular distance north of the equator in DD
7	Longitude	Angular distance east of the prime meridian in DD
8	Datum	Datum associated with the Latitude and Longitude
9	Parameter Name	Name of AQS parameter measured by the monitor
10	Date Local	Calendar date (Local Standard Time, YYYY-MM-DD)
11	Time Local	24-hour clock time (Local Standard Time, HH:MM)
12	Date GMT	Calendar date (Greenwich Mean Time, YYYY-MM-DD)
13	Time GMT	24-hour clock time (Greenwich Mean Time, HH:MM)
14	Sample Measurement	Measured value in the units specified
15	Units of Measure	Unit of measure for the parameter
16	MDL	Method Detection Limit
17	Uncertainty	Total uncertainty associated with a measurement
18	Qualifier	Indicate why values are missing
19	Method Type	Indication for the type of method used for collection
20	Method Code	Internal system code indicating the method
21	Method Name	Description of the method for collection
22	State Name	State where the monitoring site is located.
23	County Name	County where the monitoring site is located.
24	Date of Last Change	Date of Last Change

3 Analysis of Air Quality Data

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

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.
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.
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.
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.
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.
Q6.	What is the mean SO2 levels for the hottest states found in Question 6?

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

Some data may be missing or improperly formatted. It is up to you to handle such cases in your program in the manner you consider appropriate.

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 **8-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 10 points towards your final course grade. The programming component accounts for 80% of these points with the written element (to be posted later) accounting for the remaining 20%. This programming assignment will be graded for 8 points. The point distribution for this assignment is listed below.

Point Breakdown:

1 point:	Correctly configured Hadoop cluster
1 point each:	Correct answer for questions Q1-Q6
1 point:	Substantiate results in a PDF

7 Milestones:

You have 3 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-Q3 and write answers your answers to your local HDFS cluster.

Milestone 3: Complete the MapReduce implementations for Q3-Q6. Put the final touches on your report to substantiate your results from Q1-Q6.

8 What to Submit

Use the CS455 checkin program 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 description of each file and any information you feel the GTA needs to grade your program.

The folder set aside for this assignment's submission using checkin is **HW3-PC**

Filename Convention: All classes should reside in a package called **`cs455.hadoop`**. The archive file should be named as <FirstName>-<LastName>-HW<x>-PC.tar. For example, if you are Cameron Doe then the tar file should be named Cameron-Doe-HW2-PC.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	3/11/2020	First public release of the assignment.