Homework 3 - Analyzing Air Quality Data Collected across the United States using MapReduce

CS455 Introduction to Distributed Systems

Menuka Warushavithana

1. 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.

- **California**

Mapper

- Outputs the name of the state as the key, and a (custom) composite Hadoop writable consisting the county code and the site number - such that each reducer would deal with values pertaining to a single state

Reducer

- For a single state, it iterates through all county-code and site-number composite values and inserts them to a Java Set such that the set will contain only unique identifiers for monitoring sites. Outputs the name of the state (key) and the size of the aforementioned set.

2. 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.

- **East Coast**

Mapper

- If the state associated with a particular record belongs to the east coast, the mapper outputs the Text “east-coast” as the key and value of the column “Sample Measurement” (in gases), and it outputs “west-coast” if the state belongs to the west coast.

Reducer

- A single reducer receives all values for a coast (east/west), and calculates the mean of SO2 values. The reducer outputs the Text “east-coast” or “west-coast” as the key and the mean SO2 value associated with it as the value.

3. 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.

- **16.00**

Mapper

- For each record, the mapper emits the time of day (GMT) as the key and the value of the column “Sample Measurement” as the value such that a single reducer would deal with a single time of day.

Reducer

- Calculates the mean using all “Sample Measurement'' values and outputs the time of day as the key and the mean as the value.

4. 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.

- **There has been a gradual decrease in SO2 values over the last 40 years**

Mapper

- For each record, captures and emit the year as the key and the value of the column “Sample Measurement” as the value such that a single reducer would deal with values associated with a single year.

Reducer

- Calculates the mean using all “Sample Measurement” values and outputs the year as the key and mean as the value.

5. 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.

1. **Arizona**
2. **Puerto Rico**
3. **Texas**
4. **Nevada**
5. **Virgin Islands**
6. **Mississippi**
7. **Florida**
8. **Louisiana**
9. **Arkansas**
10. **Oklahoma**

Mapper

- For each record, first check if the month (GMT) is June, July, August, and then capture temperature (Sample Measurement value) and output the state name as the key and the temperature as the value, such that a single reducer would deal with a single state.

Reducer

- Calculates the mean for using temperature values and output the state name as the key and mean temperature as the value.

6. What are the mean SO2 levels for the hottest states found in Question 5?

State	Mean SO2 Level (Parts Per Billion)
Arizona	6.192774798044036
Puerto Rico	3.032422601663288
Texas	2.737068367694409
Nevada	0.6998966780914422
Virgin Islands	2.967796554157242
Mississippi	3.0614426494932
Florida	2.7050180516203124
Louisiana	3.4689261723481413
Arkansas	2.6116126071610415
Oklahoma	4.070519293754278

First, the results from Q5 are inspected manually, and then hard-coded into a list in the MapReducer job for Q6.

Mapper

- For each record, checks if the state name belongs to one of the 10 hottest (summer) states, and process only them. Capture the SO2 value (Sample Measurement value) and emit state name as key and SO2 level as value.

Reducer

- Calculates the mean by adding up all SO2 level values for a single state, and output state name as the key, and mean SO2 level as the value.