CSE5331 - DBMS Model and Implementation

REPORT - PROJECT 3 (Map/Reduce)

Project Team:

Mahesh Hingane - 1001017122

- OVERALL STATUS OF THE PROJECT -

  The project is ***completed*** and can be executed successfully. Here is the brief description of the implementation of some major components:

  1. *Class Map1* – This is the mapper class for the first part of the map/reduce task. It takes raw input files and gives the key-value pairs to the reducer for the first part. Output to reducer has stationID and year and month appended together as key. The value part is the concatenated string of temperature and windspeed values.

  2. *Class Redcuce1* – This is the reducer class for the first part of the map/reduce task. It takes input from first mapper and generated output in a file. This output extracts minimum and maximum temerature and windspeed values for each month, each year and each station.

  3. *Class Map2* – This is the mapper class for second part of the task. It takes input from first reducer and gives output to second reducer. It creates key as year value.

  4. *Class Reduce2* – This is the reduced class for the second part.It finds top 5 maximum and minimum values for the temperature and windspeed for each year. It also prints the stationid form which these values are fetched.

- FILE DESCRIPTION -

  First task of the project generated a file with maximum and minimum temperature and windspeed values for each month. The file contains staionid, year and month as key and the above mentioned maximum and minimum values as values.

  The second task generates a file with year as the key and top 5 maximum and minimum temperature and windspeed readings as values.

- LOGICAL ERRORS -

  1. Redirection of output of first task to the second task – Implementing chaining of map-reduce classes was not straight forward. I had to go through multiple tutorials to understand the syntax for this.

  2. Deciding what should be the key and value in each map-reduce task – The programming logic depends heavily on selection of key and value. So, I had to decide cautiously the keys and values.

  3. Appending the stationid to the final result – The final output required printing stationIds alongwith the temperature and windspeed readings. For this, I had to iterate over the intermediate output file twice. This may not be the best solution, but I believe it generates the correct output.

- ANALYSIS OF RESULTS -

  1. I tried to execute the code with multiple mappers and reducers, but even after modifying

the mapred-site.xml file, I couldn't get multiple reducers working. The output does not show any improvement in performance and I could not see multiple output files being generated. After some internet search, I found some links saying that this is the issue with version 2.2.0 of hadoop. This version by default supports one mapper and one reducer. With this setting, the attached program takes around ***150 seconds*** to process all 5 input files and generate the final result.