Written Question :
1.) Given that mapper and reducer function produces the correct output, what possible issue(s) could you face while processing a job consisting of about 4 million records? Suggest a workaround for that issue (without changing the number of mappers or reducers).

Answer :

The possible issue is if we divide the result based on the month, the first reducer (from A - M) will have 10 categorized in that reducer, and only two in the N-Z reducer. Therefore, with the parallelism idea that the data should be divided equally to optimize the performance, the issue could occur. The solution could be dividing the data to the first 6 months into the first reducer and the rest into the other reducer. This will make the data flow consistent.

2.) After testing on a small text file, it was noted that the pipeline does not produce correct output. Explain why this pipeline does not produce the correct output.

Answer :

In the end, there is nothing related to the sorting algorithm before returning the result of the first ten words. The reason that we have to sort is the mapReduce algorithm always shuffle the result. Therefore, by mere returning the first ten results could lead to returning results that are not the highest.