How to solve the problem

Name: Wayne

Part1: The key to solve the problem.

This is a interesting problem to improve the performance of the program, which is really similar to the problem we deal with in the cloud computing course.

I ran the original code with the sample data file (239M) , and found it was too slow to see the result. So, the first thing I did is to find the bottleneck.The time complex of most algorithm in the code is O(n).However, I found when calculate the “top 10 accepted answers”, the time complex of this part is O(n*n), because there are two nested for loops.

The general way to solve this problem is using more space to save the time. So, I use a hashSet to store the accepted answer id in single loop.
And then I search in the posts, find the answer posts, and if the set contains this answer, the count of user plus one.

In this way, I use extra O(m) space and change the time complex from O(n*n) to O(n).

Part2: Something I thought.

I also considered other loops. For example, when print the result, there are two nested for loop, the time complex is O(10n). I thought we can sort the answer map by count firstly, then get the top 10 by index. However, the time complex of sort is O(n*logn), when n is huge, the n * logn is larger than 10n. So I did not change the code of this part.

Part3: Some bugs

When I run the code , I found there is no DisplayName for some answers, I found the reason is that the User id might be null or empty, which is invalid. So I add a condition to filter the data without UserID. There are maybe other bugs when run in the real data, but I can not find it with in 2 hours.

Part4: What I will do next.

If I have more time, and the data is so huge, I will use EMR in AWS platform with many instances calculate at the same time. I can use map reduce framework.

For Map part, read the file for each line, and output is <UserId, AnswerCount>,
The input of reduce :
<UserId, AnswerCount1, AnswerCount2, AnswerCount3....>
The out put of reduce:

<UserId, TotalAnswerCount>

Then I will separate the file into several part and select the maximum value in each part ,and then select the maximum from the result.

For example:
In File1:
UserId count
User1 10
User2 20----------> User2 20----------|
In File2: |
User3 5 |
User4 30----------> User4 30----------> maximum is User4 30

Then delete the user4 in the file, repeat 10 times and get the top 10.
Similarly, we can use this way to find the top 10 accepted answers.