1. What is the key need for this algorithm/approach/system? What are some short comings of existing approaches?

The map step and the reduce step are the key element to accomplish this algorithm(MapReduce). If we want to handle streaming data, it might have problems because we do not have all the data on hand. MapReduce is the batch process which deals with huge amounts of data which we have. However, the problems MapReduce solved are much more than the problems it causes. We cannot also expect the result will be produced in seconds. Therefore, I do not think the speed problem is the shortcomings for MapRdeuce, but it depends on the usages I would like to have.

2. What is the key object (term) that is the solution to this need? (e.g. tweet, RDD, Tensor) Describe this object in a few sentences. This is the fundamental concept or "thing" proposed to solve the need addressed in question #1.

The map step converts a set of data into tuples (key/value pairs). After map step, the reduce step will take the date from map step as input and combine those tuples into a smaller set of tuples as output. By doing so, the size of data will be highly reduced and these files are partitioned into multiple files. Hence, they can be used in the distributed applications.

3. What has the author identified as a weakness or limitation of the proposed algorithm / approach / system? Or what has the author proposed as next steps? If the author does not provide this information, what do you think could be improved?

MapReduce perform batch data process, so it is not good to use in stream data processing. Moreover, the Map step and Reduce step will break down the data into key value pair and then reduce step produces the final output. Nonetheless, these two steps would require lots of time to perform tasks, so they will increase latency.

4. What is something interesting your learned from this paper, or your thoughts about its strengths and/or weaknesses. Is there anything else interesting about this paper, or your interpretation of it that you want to share?

In this paper, the author explained how to implement distributed application using MapReduce method. This is very important for me because I do not know how to implement parallel and distributed system. However, by using MapReduce, I can just take the partitioned files and put them into multiple machines. Moreover, the way they deal with fault is attracted to me since in real life, it is a very difficult problem. Because

our data sets are run on hundreds of machines, when certain reduced task fail, we might need to execute map step again. However, if the reduced task completed, the output will send back to global file system. Therefore, we do not need to do all the process again because of fault, but just redo the step we fail.

5. Read an additional related paper. Provide a citation for this paper. How does this paper relate to the assigned one? What new information did you learn by reading it? This answer should be a decent sized paragraph describing the content of this paper (be specific) and how it relates to the assigned paper.

Ref:

Maitrey S, Jha CK. MapReduce: Simplified data analysis of Big data. Elsevier. 2015; 57:563–71.

This paper focus on Hadoop and its implementation of MapReduce for analytical processing. By reading this paper, I understand how to analyze the large amount of multi-structured data sets. It compared the difference between the standard SQL which are employed by relational database systems and the procedural nature of MapReduce. It also taught me how to deploy Hadoop to implement MapReduce. The previous paper provided me with the basic theory of MapReduce. This reading helps me have the whole picture regarding how to utilize the MapReduce in the real world.