**CSP554—Big Data Technologies**

**Assignment #1 (Modules 01a & 01b)**

**Summary of the article “The parable of google flu”: (Question-4)**

This article discusses the mistakes that Google Flu Trends (GFT), made in making predictions of Flu affecting and the rate at which it was spreading. Here, the author didn’t try to advocate that the traditional methods which were being used by Centre for Disease control and Prevention (CDC) were better in making predictions since they were giving much accurate results, since it had its own problem like the amount of time spent in prediction in comparison to GFT. Rather, the author is trying to make a statement that a new approach must be taken where both the methods, new ones like Big Data algorithms and the old traditional approach with hypotheses. As we can see in the article, though the rate at which GFT is able to process data and bring out results much faster than the CDC, but the error rate is also equally higher than that of CDC. But when the GFT and CDC methods were combined, the error rate was the lowest of the three. The author also explains as where the error is occurring in GFT. There are two issues: one is Big Data Hubris where there was a belief that Big Data are a big time replacement for conventional approach. But the evidence states that has turned out to be a misconception since the size of the data cannot make the core problems of measuring and constructing validity with reliabilities and dependencies among data. There were many instances which showcased that overfitting was happening which in turn was affecting the accuracy. Because of this GFT was not able to see the coming 2009 Influenza A-H1N1 at all. Even though the system was updated after this, it still had an issue in finding useful search terms among the lots. The second issue was algorithm dynamics, where the developers were constantly changing the workings of the system to meet their business ends. The author agrees Google is a company which needs to earn revenue but when it is done at the cost of its accuracy, then the system will fail to serve its purpose. The author later on discusses about the “blue team” dynamics where the system the algorithm producing the data has been modified by the service provider and the “red team” dynamics where research subjects attempt to manipulate data to meet their business end. The latter one is the most effective one where even the slightest manipulation can cost a political leader his seat or a product kill its competitors.

The author has correctly expressed the problems and concern of this new age technology where information is becoming a valuable asset. I agree that we must put emphasis on combine both old and new approach to create an “all data revolution” (as the author suggests) which will in turn help in creating a much deeper understanding of the problems that is being analysed.

**Summary of the paper “Faults are not just crashes”: (Question-6)**

In this paper, the authors are trying to present a solution which can help us make a Hadoop system that will be Byzantine fault tolerant along with being a crash tolerant. The existing systems are surely crash tolerant since they can be easily detected (by using checksums) but that is the case with arbitrary faults since they are silent (thus not getting detected) and corrupting the results. There were various solutions for this one of them being making MapReduce execute each job twice and re-execute if the results do not match. But this is not at all cost -efficient.

The solution which the paper presents tries to do it efficiently. Though the authors confess that this new system can take more time and resource to execute a task in comparison to the original one, but not every task requires such system. Only the applications which require a high degree of correctness in their results would want to use it. The algorithm is divided into various executions like “deferred execution” in which the JobTracker starts only f+1 replicas of the same task reduce them to see if they get the same results. If they don’t get the exact results, then only more replicas are made. Then there is another execution “Tentative reduce execution” which tries to reduce end-to-end latency for the job completion. Next one is “Digital outputs” where ‘hashes’ are used since the results are large, it will be stored as digests to validate the output of other results.

In my opinion, this solution is quite good in comparison to the naïve solution, as this would be cost-efficient. But as we can see in the experimental evaluation section, its resource consumption and the duration for doing splits is quite large when compared to the original model. So, this solution does not have wide range of applications. This can be used in fields where scientific computations are done, like in Astronomy, particle exploration also in banking sectors and security. There may be more applications but using it for problems with requirement of less degree of correctness will give results but won’t be efficient in that situation.