**ASSIGNMENT - 1**

**CSP - 554 BIG DATA TECHNOLOGIES**

**QUESTION #3**

* No, I don’t have practical experience on data mining, machine learning, statistics, data science and big data but I have basic theoretical idea about the afore said topics. In this Spring I have taken Big Data Technology and Data Mining as well to gain knowledge and I have planned to take Machine Learning in the upcoming Fall semester. In future I want to do a lot of projects by relating all topics and I want to make my career in Big Data.

* I always feel that it would be helpful to learn any technology from the scratch as it is easy to understand the basic concepts. I have started learning about many components of Hadoop from Udemy course but due to tight schedule for first semester I have not able to explore a lot. I am very much glad to have an opportunity here to learn new things about Hadoop and I have plan to go through all components of Hadoop within this semester so that I can gain a solid knowledge over it. I want to learn Scala and Spark as well.

* The additional topics which are in the scope of the course and are of special interest to me are Zoo keeper, HBase and Impala.
* While learning Big Data course from Udemy I gained knowledge with the few components of Hadoop like Map Reduce, Pig, Hive, Sqoop. To use these components, I have installed the Hadoop ecosystem on the Virtual Machine in my Laptop.
* No, I do not have any such personal issues and I have waited from late November to get into this course as I really want to take this course in this semester. It is my privilege to attend the lectures and to gain knowledge from there.

**QUESTION #5**

**The Parable of Google Flu: Traps in Big Data Analysis**

The prediction made by Google in estimating the Flu cases which is popularly known as Google Flu Trends, is based on search terms and is considered as not that profitable as it was expected. According to Author following are the reasons:

1. Big Data hubris
2. Algorithm Dynamics

Big Data hubris assumes that all the traditional data collection and processing techniques are being replaced by this new technology i.e. “Big Data”. Here in this article the Author inspects this theory by stating that however this technology has its wide advantages, but it cannot be acted as a substitute to the conventional collection and analysis data techniques. The factors that support the traditional techniques are validity, reliability and dependency.

Initially the GFT searches the best match from the numerous search terms that best fits the data points. In due course this turned out to be unsuccessful because all the terms that matches could not effectively portray the occurrence of Flu. It supported the statement that considering Big Data was over fitting many small cases, which is the primary concern in conventional analysis techniques.

In current week, the trends which are being forecasted are relative to the results that are posted during previous week. These patterns define that GFT omits the standard information that can be mined using the conventional statistical methods. According to Author Algorithm dynamics is considered as another factor. Many changes are made in search algorithm to enhance the customer satisfaction affected the tracking capacity of GFT resulting as drawback to Google Flu. The factors that are important when dealing with Big Data are "Transparency" and "Replicability" of the data, and some essential characteristics are using the Big Data to understand the unknown, Studying the algorithms and the fact that just don't consider the size of the data, but also indulge the traditional methods to extract the unknown information.

According to Author, Big Data hubris and the Algorithm dynamics are the only reasons for the failure of GST, but I feel apart from these two factors there might be many other factors like demographics, data failure, search for reference or scientific purpose could be the reasons for failure. The Author may be right in his own place but before concluding he should consider about other possibilities also.

**QUESTION #7**

**Byzantine Fault Tolerant MapReduce**

To run critical jobs like scientific data analysis MapReduce is used but arbitrary faults do occur and can probably corrupt the results of MapReduce jobs. MapReduce runtimes like Hadoop tolerate crash faults, but not arbitrary or Byzantine faults. According to the authors, BFT MapReduce uses twice the resources of the original Hadoop, instead of the 3 or 4 times more that would be achieved with the direct application of common Byzantine fault-tolerance paradigms. This cost is acceptable for critical applications that require that level of fault tolerance.

The authors claim that the main contribution of the paper is an algorithm to execute MapReduce jobs tolerating arbitrary faults. They implemented the BFT MapReduce by modifying Hadoop and measured its performance using the Hadoop’s Gridmix benchmark.3 They say that the results confirmed that indeed it is possible to run a Byzantine fault-tolerant Hadoop using twice the resources of Hadoop.

I feel that the BFT MapReduce Algorithm is a excellent addition which could be used for tackling byzantine faults. However, the cost of execution remains expensive which is why I feel that we should be using it only when we need fault tolerance at the highest level.