**Ritika Kumari**

**A20414073**

**CSP #554**

**Assignment #1**

1. Submit very brief answers (or bullet points) to the following questions:

* Describe any prior experience you might with, data mining, machine learning, statistics, data science and big data

**Answer:** In my first semester I took Data mining which involves finding interesting patterns from datasets and used R language to perform explanatory analysis. Also, in previous work experience, I worked for big data technology for 1 year, So I have basic idea of Big data. I also worked on java and PL/SQL technology for 3 years.

* Share any big data interests and personal learning goals for the course.

**Answer:** Since Big data involves large scale storage and processing of large data sets. So, data mining of big data is very interesting and is getting lot of attention currently. I am expecting that after the completion of the course I will be capable of applying my data mining logic in big data and want to be an expert NoSQL Database.

* Indicate if there are additional topics in the scope of the course of special interest to you.

**Answer:** I know SQL very well but most of the database types are represented by NoSQL database. Considering all the advantages of NoSQL such as better performance, adaptability, scalability etc., it would be great to learn NoSQL in depth.

* Indicate if you have access to big data technology and data sets, of what nature, and in what industry.

**Answer:** Yes, I do have experience in this field where I had access to the big data technology and data sets in software industry. I was responsible for pulling the data from various structured and unstructured data sources into Hadoop platform and standardize all data through a series of master data management processes to reach client goal to become efficient and effective within a different geographic location in attracting more customers and improve their user experience.

* Do you have any anticipated personal issues such as expected absences or other necessary accommodations with course impact? (Of course, these will be held in strictest confidence.

**Answer:** No

1. Read article on “Blackboard” in Articles section

* The Parable of Google Flu (just 3 pages!)

1. Summarize the main points of the above article and your thoughts (questions you might want to ask the authors, areas where you disagree, other comments)

* No more than about ½ page single spaced
* Submit via blackboard

**Answer:** This paper basically talks about GFT’s failure and all the possible factors contributing to the failure. GFT overestimated the pervasiveness of flu in the 2012-2013 season, and indicated almost the double of the genuine levels of flu in 2011-2012. On the top of this, from August 2011 to September 2013, GFT over-anticipated the commonness of flu in 100 out of 108 weeks. It states that many sources of 'big data' come from private companies, who, just like Google, are constantly changing their service in accordance with their business model. Two issues that contributed to GFT’s mistakes— big data hubris and algorithm dynamics. “Big data hubris” is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis. Algorithm dynamics refers to the continuous changes made by developer for improving search algorithm of google search. Developers made some changes in search algorithm in 2009 after which H1N1 flu did not got detected by GFT.

After reading the article completely, the first thing which came to my mind is numbers and information can be basic devices in bringing complex issues into centre. The comprehension of sicknesses, for instance, profits by calculations that assist monitor their span. Yet, without setting, a number might be only a number, or even misleading. Google Flu Trend was designed to provide real-time monitoring of flu cases around the world, however it additionally represents where 'enormous information' examination can turn out badly.

We require a superior comprehension of how this influences the information they create; else we risk reaching erroneous inferences and embracing inappropriate arrangements.

1. Read article on “Blackboard” in Articles section

* Byzantine Fault Tolerant MapReduce

1. Summarize the main points of the above article and your thoughts (questions you might want to ask the authors, areas where you disagree, other comments)

* No more than about ½ page single spaced
* Submit via blackboard

**Answer:** The fault tolerance mechanisms of the original MapReduce cannot deal with Byzantine faults. These faults in general cannot be distinguished, so they can silently corrupt the output of any map or reduce task. Byzantine Fault Tolerant MapReduce algorithm hides these faults by executing each task more than once, differentiating the outputs of these executions, and dismissing non-matching outputs.

Byzantine Fault-Tolerant MapReduce focuses three major problems in Hadoop environment:

i). Confining the programming model makes it simple to parallelize and appropriate calculations and to make such calculations fault tolerant.

ii). Various enhancements in proposed framework are in this manner focused at decreasing the measure of information sent over the system: the locality optimization enables us to read information from local disks and composing an only copy of the intermediate data to local disk spares cost.

iii). Surplus execution can be utilized to lessen the effect of moderate machines, and to deal with arbitrary failures and data loss.

The paper introduces a Byzantine blame tolerant MapReduce calculation and its trial assessment. The test assessment affirmed with f = 1 the assets utilized and the make span basically twofold. Although this is an extensive cost, we have demonstrated that it is vastly improved than elective arrangements, state machine replication and result examination conspire.