**CSP 554 – Big Data Technologies**

**Adarsh Mathad Vijayakumar – A20424847**

**avijayakumar@hawk.iit.edu**

3) (5 points) 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

In my previous semester (1st semester) I had taken Data Mining course, where in I learnt about the key concepts like Decision trees, Association rules, Map Reduce, Recommendation Systems and many other related to the field. I also learnt about the Weka tool wherein I got to learn few of these concepts in practical by working on some real datasets.

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

I am interested in learning about the latest big data technologies such as Hadoop, Kafka, Zookeeper, Hive, Pig, Spark, MongoDB. I am hoping to get hands-on experience on few of these as well.

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

From the course syllabus point of view, I feel that it pretty much covers everything that I’m interested in.

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

Presently I do not have any access.

* 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.)

None

4) Read article on “Blackboard” in Articles section

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

(5 points) 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)

The article “The Parable of Google Flu: Traps in Big Data Analysis” depicts how Google Flu Trends (GFT) unpleasantly failed on influenza prediction. GFT was built to predict Centers for Disease Control and Prevention (CDC) reports and was over estimating than the CDC estimates. The paper also explores two primary issues that contributed to GFT’s mistakes that is, big data hubris and algorithm dynamics, and offer lessons for moving forward in the big data age.

It talks about the big data hubris where it is often assumed that big data is substitute for rather than supplement to traditional data collection and analysis. One of the noteworthy example for the big data hubris is the overestimation of influenza cases by GFT. This does not imply that Big Data Analytics are futile. The author means to say that such analytics play more of a supportive role than a substitutive role to traditional means of data collection and analysis. The cause of this issue could be because GFT overlooked considerable information which could have been easily extracted by traditional statistical methods.

Algorithm dynamics are the changes made by engineers to improve the commercial service and by consumers in using that service. The flawed algorithm dynamics may have caused Google to misrepresent the reality of flu trends. The article highlighted how this problem is not pertaining to Google alone, even social networking sites like Facebook and Twitter have been projecting wrong results due to external contributions and hence this is a problem which needs to be tackled by everyone dealing with data who are looking to get maximum benefit from it. Probably it requires a change in attitude on how to deal with data, and realizing big data is not here to replace traditional statistical methods rather to work hand-in-hand.

I agree with the author’s sentiments and feel that new technologies are not born to replace old, rather we need to use the best of all worlds and get the maximum benefit. Our end goal is to use the data we have to solve problems and move forward in the big data age.

Extra Credit:

6) Read article on “Blackboard” in Articles section

* Byzantine Fault Tolerant MapReduce

(2 points) 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

The article talks about basic introduction to MapReduce and Hadoop and how Map Reduce was designed to be fault tolerant because of the scale of computers, switches, routers and power units involved. Then the article talks about the specific fault-tolerant system called Byzantine fault-tolerant (BFT) and its advantages and challenges. The article also states the BFT algorithm and prototype as well as its evaluation using Hadoop’s Gridmix benchmark in the Grid’5000 testbed. It also confirms the evaluation and what might be intuited from the algorithm that when that with f = 1 the time to execute a job essentially doubles in a small cluster, which is

equivalent to approximately the double CPU time. Author argues that when f = 1 is a realistic assumption because: (i) arbitrary faults are rare (ii) it means that the probability of more than one faulty replicas of the same task returning the same output is negligible.

The article has also explained the concept of the data restoration where in case if the data is corrupted, the same can be restored with checksum. There are many ways in which the fault may occur, like from the crashes of map and data corruption on disks and many other ways. The article also states that BFT MapReduce tolerates any number of faulty task executions at a low cost: the re-execution of that task. But, this is not what happens with a simplistic solution like executing a job more than once using the original Hadoop and comparing the outputs. If each execution was affected by one fault in any task, the job might be re-executed forever without any two outputs ever matching.