**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
  + Dotnet developer with 4 years of experience in web development
  + Knowledge in Javascript, Spring , Hibernate, data mining technique and azure
  + Hands on experience on R to perform explanatory analysis
  + Hands on experience on PL/SQL.
  + Understanding on LINUX environment.
* Share any big data interests and personal learning goals for the course
  + I want to learn various Hadoop component like Hive, Pig, stream processing and Spark
  + Understanding practical implementation of Big Data Technologies like Hadoop
  + Exploring NoSQL Database and getting deep knowledge of it.
* Indicate if there are additional topics in the scope of the course of special interest to you
  + Practical knowledge of various Hadoop concepts such a YARN,HIVE, Hadoop I/O, Zookeeper .
  + In depth knowledge of DFS, Hadoop concept, MapReduce, NoSQL Db.
* Indicate if you have access to big data technology and data sets, of what nature, and in what industry.
* Yes, I have experience of 4 year in this field where I used to work on dataset retrieved using big data technology in software industry.
* I used to get dataset from various type of structured and unstructured data source into Hadoop Platform. Then I used to manipulate data to show required field and give only required data to the client
* 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.)
* No, I don’t have any personal issues

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)

**Ans :** Author tells about Google Flu Trend (GFT) which came to headline in 2013 because of wrong prediction of ILI that it made. Two main reason that lead to GFT’s mistake was Big Data Hubris and Algorithm dynamics .According to Big Data Hubris ,Big data can substitute traditional data collection and analysis. However ,in 2009 with H1N1 flu outbreak ,GFT failed to detect it as it had keyword ” flu detector” and “winter detector” for flu.However,H1N1 was not winter flu .Since then majority of time, GFT Engineers have missed flu season with considerable margin with predictable errors each time. However, when we merge lagged CDC data with GFT, the performance is better than CDC and GFT alone .Here, I think the source of data that GFT was using ,didn’t had reliable information. For instance, If a

person search for FLU in google ,it does not reflect that the person is suffering from FLU. GFT should have considered other source of data apart from google search . Another reason for GFT’s mistake was Algorithm Dynamics i.e. the continuous changes done by google in the search algorithm to provide customer better service. Google did not documented 45 google search word used for GFT Algorithm. Also, “Blue team” changes done in 2011-2012, adversely affected GFT estimates. Re-engineering and undocumented previous version also made it difficult to understand GFT algorithm’s failure reason. Apart from blue team issue, ”Red Team” attack also affect the data generating process and keeping actual information from the people. According to Author, GFT is one of the important case study as it made us learn few very important lessons that need to be considered in Big data industry. Some of them are Transparency and Replicability, implementing the algorithm considering future scenario, documenting the work done and utilizing big data to know the unknowns.

Author concluded on a clear note which is true that “Instead of focusing on a “big data revolution,” perhaps it is time we were focused on an “all data revolution,” where we recognize that the critical change in the world has been innovative analytics, using data from all traditional and new sources, and providing a deeper, clearer understanding of our world”.

3.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)

**Ans :** MapReduce, a framework developed by google, processes large datasets. Map Reduce receives messages in the form of heartbeat. Map Reduce has some level of fault tolerance as it can detect crash when heartbeat does not come back. However, it cannot detect arbitrary faults. Byzantine Fault Tolerance was introduced to detect these arbitrary faults. Byzantine Fault Tolerance algorithm creates 2f+1 replicas of the data where f is the maximum fault that can occur. However, arbitrary fault are uncommon. So, only f+1 replicas are made instead of 2f+1 replicas. Once, these replicas are created, they are compared with each other and if all the replicas are same then there is no arbitrary fault in the data. However, if replicas doesn’t match, then there is some arbitrary fault. While experimental evaluation of this algorithm, the performance of algorithm was analyzed. The output came same as expected. However, since the data increases, sometimes map task cannot get executed where input split is stored. In that case, the map task gets to take data from network, the beginning of next task gets affected. So, performance decreases. Hence, BFT version had better localy than the official version.BFT MapReduce tolerates any number of faulty task executions at a low cost : the re-execution of that task.