**CS595 - Assignment 1**

1. ***Submit very brief answers (or bullet points) to the following questions:***
   1. ***Describe any prior experience you might with, data mining, machine learning, statistics, data science and big data***

**Answer:**

* Data Warehousing professional with 6+ years of experience in BIDW domain.
* Knowledgeable in statistical and data mining techniques and algorithms including but not limited to Classification, Clustering, Association Rules, Anomaly detection, Logistic Regression
* Hands on experience in Python and R to perform exploratory analysis on data
* Technical expertise in ETL Tool Ab Initio(Co>operating System: 3.1.2.5, GDE version: 3.1.1)
* Strong Command on UNIX and PL/SQL. Worked on Teradata, Oracle, and DB2 databases
* Worked with HDFS file system for creating Data Lakes
* Relevant Courses at IIT: Data Mining, Data Integration, Warehousing and Provenance, Data Preparation & Analysis, Advanced Database Organization.
  1. ***Share any big data interests and personal learning goals for the course***

**Answer:**

* Gather substantial knowledge about Big Data Technologies (PySpark, Hive)
* Gather in-depth knowledge of Hadoop framework
* Learning Map-Reduce, NoSQL Databases
  1. ***Indicate if there are additional topics in the scope of the course of special interest to you***

**Answer:**

* Hand’s on experience of Hadoop components such as Hive, Pig and Pyspark
* Hand’s on experience of Map-Reduce, NoSQL Databases
  1. ***Indicate if you have access to big data technology and data sets, of what nature, and in what industry.***

**Answer:**

* During my professional tenure, I have had access to HDFS and datasets of our client Allstate Insurance dealing with Insurance (Claim, Policy, Billing), geographical, demographic & Clickstream data.
  1. ***Do you have any anticipated personal issues such expected absences or other necessary accommodations with course impact? (Of course, these will be held in strictest confidence.)***

**Answer:**

* None

1. ***Summarize the main points of* “The Parable of Google Flu” *article and your thoughts (questions you might want to ask the authors, areas where you disagree, other comments)***

**Answer:** Following is the summary of “The Parable of Google Flu” article:

* The article talks about Google Flu Trends (GFT), which is a prediction system for the tracking of the Flu using Big data technologies. The authors in this article try to highlight two key issues that contributed to GFT’s mistakes:

1. Big data hubris and
2. Algorithm dynamics

and offer important lessons learned from these failures for future Big data analysis applications reference.

* **Big Data Hubris:** The term “Big data hubris” often refers to implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis, i.e. in other words, quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data.
* The authors highlight one of the core challenge that most big data applications that have received popular attention lack of instruments to produce valid and reliable data amenable for scientific analysis.
* As in the case of GFT, the developers weeded out search terms that matched the propensity of the flu but were structurally unrelated, and so did not help predict the future in adhoc manner, but in reality, this caused some of the important data points such as seasonal search terms unrelated to the flu but strongly correlated to the CDC data to be left out and caused prediction model overfitting.
* By the example of GFT, the authors stress the fact that big data is no substitute for information that could be extracted by traditional statistical methods and ongoing evaluation and improvement, but, by incorporating this information, we can make big data application systems more efficient and accurate, i.e. Big data to be used for Information Augmentation(IA) rather than purely for Artificial Intelligence(AI).
* **Algorithm Dynamics:** Algorithm dynamics are the changes made by engineers to improve the commercial service and by consumers in using that service. In the case of GFT, the culprit is changes made by Google’s search algorithm itself.
* The authors talk about the phenomenon of “Blue team” dynamics — where the algorithm producing the data (and thus user utilization) has been modified by the service provider in accordance with their business model. The article also introduces another similar phenomenon, Red team dynamics that occur when research subjects (in this case We searchers) attempt to manipulate the data generating process to meet their own goals, such as economic or political gain.
* By highlighting Algorithm Dynamics issue for GFT case study, the authors warn big data engineers that the more successful we become at monitoring the behavior of people using these open sources of information, the more tempting it is to manipulate those signals for economic or personal gains.
* Finally, the authors point out the critical lessons that can be learnt from the GFT case study as we move forward in the age of big data analysis.
  + 1. **Transparency and Replicability:** Follow standards and publish/document papers related to big data analytics application for future maintenance and reference standpoint.

* + 1. **Use Big Data to Understand the Unknown:** To explore and gather the data for problem at hand at the very granular level as possible for big data analysis to predict accurate results.
    2. **Study the Algorithm:** Researchers/Engineers need to have a through knowledge of the algorithm that is being implemented and need to ensure robust patterns backed up with statistical analysis before making subsequent changes to the algorithms.
    3. **It’s Not Just About Size of the Data:** We should aim at collaborating both traditional applied statistics with big data, as traditional “small data” often offer information that is not contained (or containable) in big data, and the very factors that have enabled big data are enabling more traditional data collection. Instead of focusing on a “big data revolution,” perhaps we should focus 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.

***Extra Credit:***

1. ***Summarize the main points of* “Byzantine Fault Tolerant MapReduce” *article and your thoughts (questions you might want to ask the authors, areas where you disagree, other comments)***

**Answer:** Following is the summary of “Byzantine Fault Tolerant MapReduce” article:

* Even though the MapReduce runtimes like Hadoop can tolerate crash faults, but they do not tolerate arbitrary or Byzantine faults. The number of failures are frequent, so MapReduce was designed to be fault tolerant for crashes. However, it is incapable of handling arbitrary faults that can lead to corrupt or incorrect results of MapReduce jobs.
* One solution to deal with arbitrary faults is Byzantine Fault-Tolerant(BFT) approach and this works for a service that uses client server model and the other solution is to execute each job twice and check if the results do match.
* In the BFT MapReduce algorithm, the job tracker starts with f+1 replicas of each map task and assigns to each task tracker which sends heartbeat messages in return. If job tracker notices that there are no f+1 replicas that match then it starts another replica of same map.
* If the replicas match the job tracker inserts them in the queue. Then reduce task happens and stores the output and Hash in HDFS. Hadoop is written in java and Mostly, modifications are made in “Job Tracker” Component.
* We perform the experiments in Grid’5000. The experiment says that when f=1 the time to execute a job doubles in small cluster which is equal to double the amount of CPU time approximately. Most importantly, MapReduce handles many number of faulty task executions at low cost But, this doesn’t happen in original Hadoop.