CSP554—Big Data Technologies

Submitted by - Saptarshi Chatterjee, A20413922

Assignment #1 (Modules 01a & 01b, 10 points + 2 points extra credit)

1. (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

Took Data Mining Class last semester under Prof V. Gurbani and Machine learning Class under Prof Mustafa Bilgic this semester. Worked with Big Data and Data Science before mostly on proprietary tools at various organization.

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

My career goal is to become Big data Architect at fortune 500 company or a promising startup. Hopefully this course will be a building block to achieve that goal. Want to be conversant with all the modern big data technologies like Spark, kafka , storm , cassandra etc.

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

Want to do a deep dive on MLlib.

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

I am one of the developer of a platform named nuggetai.com . It tracks user events (writing speed , time spends, spelling error, no of spelling edits) while writing an essay / case study , and tries to classify candidates and their proficiency based on the event data using machine learning.

Available data is relatively small now , but it's growing .

Industry - Education / Hiring .

   - 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.)

Nothing as of now.

2. Read article on "Blackboard"
   - 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)
   - No more than about ½ page single spaced
   - Submit via blackboard

This article is an allegory on Google Flu Trend (GFT) which made headline in 2013 because of inaccuracy of of it's prediction on influenza-like illness. Less than optimal outcome of GFT project can be attributed to Algorithm dynamics and Big Data Hubris . 'Big Data Hubris' refers to an assumption that Big data can substitute traditional data collection and analysis. However ,in 2009 with H1N1 flu outbreak ,GFT failed to detect it because the flu was nonseasonal . 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 .

In my opinion, the source of data that GFT was using ,didn't had reliable information. 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 search keywords.

Another moving part in GFT project, was Algorithm Dynamics i.e. the continuous changes done by google in the search algorithm to provide customer better service. Apart from "blue team issue" , "Red Team" attack also affect the data generating process and keeping actual information from the people.

This article touches upon important lessons that need to be considered in Big data industry. Some of them are Transparency and reproducibility, implementing the algorithm considering future scenario, documenting the work done and utilizing big data to explore the unknowns.

Authors 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".

Extra Credit:

3. Read article Byzantine Fault Tolerant MapReduce

4. (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

This academic paper by a group of researchers from Technical University of Lisbon proposes a solution to efficiently eliminate errors due to random faults (Byzantine Fault) in a Map-Reduce system .

Hadoop inherently provides fault tolerance for system crashes and system unavailability errors , but it doesn't provide any solution if a node in-correctly computes the Map or Reduce tasks due to DRAM error , CPU or Chipset faults etc.

The core idea behind the approach is - Duplicating each map and reduce tasks into 2 copies (To be accurate f + 1 copies , where f is number of faulty replicas that return the same output . It's safe to

assume f ==1) . Then creating a digest of the result output and checking if the digest are same , across replicas , to ensure the result is not corrupted by Byzantine Fault . If the digests doesn't match the proposed system discards the result and re-executes the task until f+1 outputs match.

The proposed approach is a drastic improvement over a simplistic approach of executing 2f+1 replicas of each task and picking most voted results or state machine approach that requires 3f+1 replicas . Although proposed solution twice more expensive than using the original MapReduce runtime, however, this cost is acceptable for critical applications that require a high degree of certainty.

Though this system assumes clients and JobTracker are always correct which might not always be true and may lead to error, but I think it's a great design for systems that require extremely high degree of computation correctness.