Big Data Technologies [CSP 554 - 01]

Assignment-1 [10 points]

Anusha Sonte Parameshwar [A20518694]

[1 point]

1. What was the problem with the Google flu detection algorithm?

The problem with the Google flu detection algorithm was it was predicting more than double the doctor visits for influenza like illness than the Centers for disease control and prevention (CDC).

Less than optimal outcome of GFT can be attributed to 'Big data hubris' and 'Algorithm dynamics'.

[1 point]

2. What is Big data hubris?

Big data Hubris refers to an assumption that big data can substitute for, rather than be a supplement to, traditional data collection and analysis.

The main challenge lies in the fact that many high-profile big data sets are not the result of instruments designed to generate valid and reliable data suitable for scientific scrutiny.

In GFT's context, the source of data used by GFT did not have reliable information. For example, if a person searches for 'FLU' in google, it does not necessarily mean that person is suffering from FLU.

[1 point]

3. What approach could have been used to improve the Google flu detection algorithm?

To improve the GFT algorithm:

- Use big data as a supplement and not as a substitute.
- Dynamically recalibrate the GFT.
- Combining GFT with lagged CDC data.
- Considering real time data for predictions.
- Avoiding the blue team and red team dynamics for data gathering.

[1 point]

4. What is “algorithm dynamics?”

Algorithm dynamics refer to the changes in the behavior of algorithms over time under various conditions. The behavior can be influenced by environmental changes, input data or changes to the algorithm itself.

So, it is important to consider how the algorithm behaves in various conditions to avoid unexpected or undesirable outcomes.

In GFT’s context, several changes to the google search algorithm and the user behavior likely affected the GFT tracking.

[1 point]

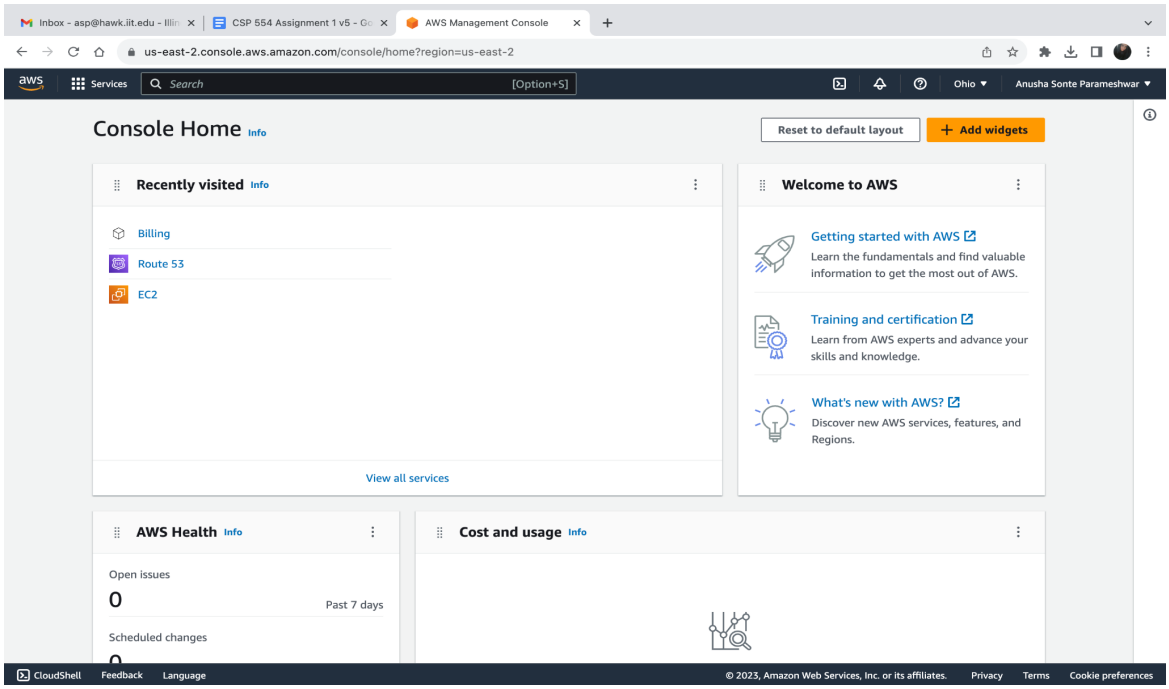
5. What aspect of algorithm dynamics impacted the Google flu detection algorithm?

Google's search algorithm is not static; more than 86 changes were reported in a span of two months in 2012. This dynamics to improve the customer satisfaction and commercial service impacted the GFT algorithm.

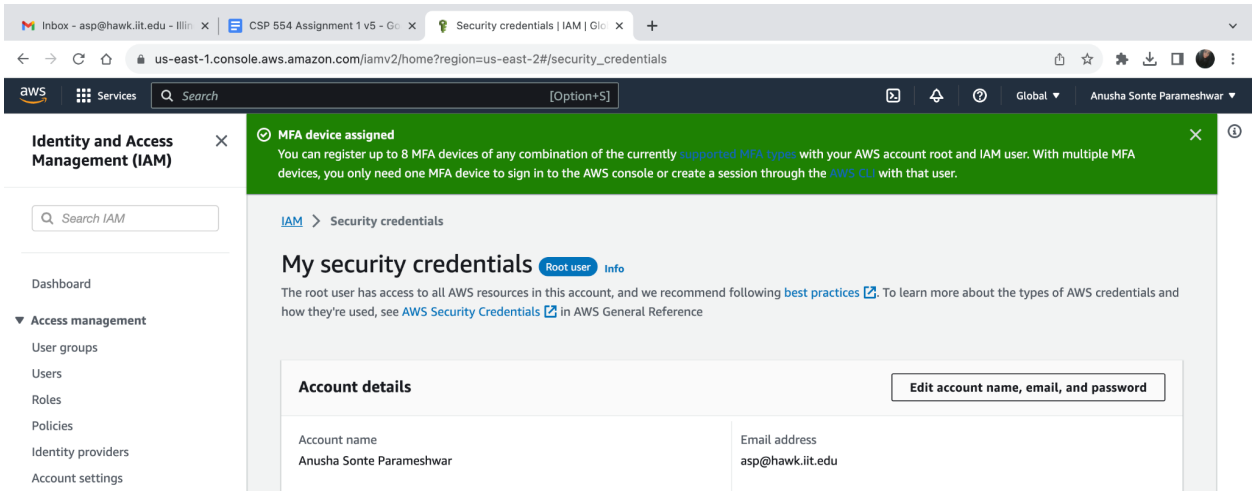
Apart from that, few other issues which impacted the GFT algorithm are as below:

- Recommended searches based on what others have searched will increase the relative magnitude of certain searches.
- Google correlate, which allows users to identify the search data is limited to national level, whereas GFT was using regional level data.
- Blue team dynamics, where the algorithm producing the data has been modified as per business model also affects the GFT predictions.
- Red team dynamics where the data generation process is manipulated for economic or political gain.
- Not considering the endogenously cultivated behavior during search.

[3 points]
AWS Account Creation:



MFA Configuration:



[2 points]
Bucket Creation:

Uploaded ‘Big Data Course Information.docx’ file into the bucket ‘a20518694csp554’ :