**Name:** Deep Pawar(A20545137)
**Professor:** Joseph Rosen
**Institute:** Illinois Institute of Technology

# CSP 554: Big Data Technologies

Fall 2024 - Assignment 1

- **Questions and Answers:**

4. (5 points) Answer each of the following questions about the article in just one to three sentences each:

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

**Ans:**

There are two problems with Google's flu detection algorithm which are algorithm dynamics and big data hubris. The primary issues were "algorithm dynamics," which showed how modifications to Google's search algorithm and user behavior impacted search patterns, which over time resulted in persistent errors, and "big data hubris," which is the belief that large amounts of data might replace traditional measurement methods. This causes GFT's initial version to overfit since it only matched search phrases with a small amount of CDC data points.

ii. What is big data hubris?

**Ans:**

Big data hubris is a mistaken belief that traditional collection and analysis data techniques can be replaced by vast quantities of data, rather than being used alongside them. This assumption frequently causes basic problems with data analysis such as data dependencies, construct reliability, and measurement validity. In the instance of GFT, this led to an overestimation of the predictive power of search term correlations.

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

**Ans:**

Google flu detection algorithm can be improved in the following ways:

- Integrating GFT data with other sources of near-real-time health data, especially lagged CDC data.
- Adapting the GFT algorithm dynamically over time to take into consideration modifications in search patterns and Google's search algorithm.
- Analyzing and making adjustments for algorithm dynamics in a way that modifications to Google's search algorithm and user behavior impact search trends over time.

iv. What is “algorithm dynamics?”

**Ans:**

Algorithm dynamics is the term used to describe how engineers modify a commercial service, like Google's search algorithm, and how these modifications lead to changes in user behavior. These dynamics change the process of generating data, which might impact the stability and accuracy of data-driven models like Google Flu Trends (GFT).

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

**Ans:**

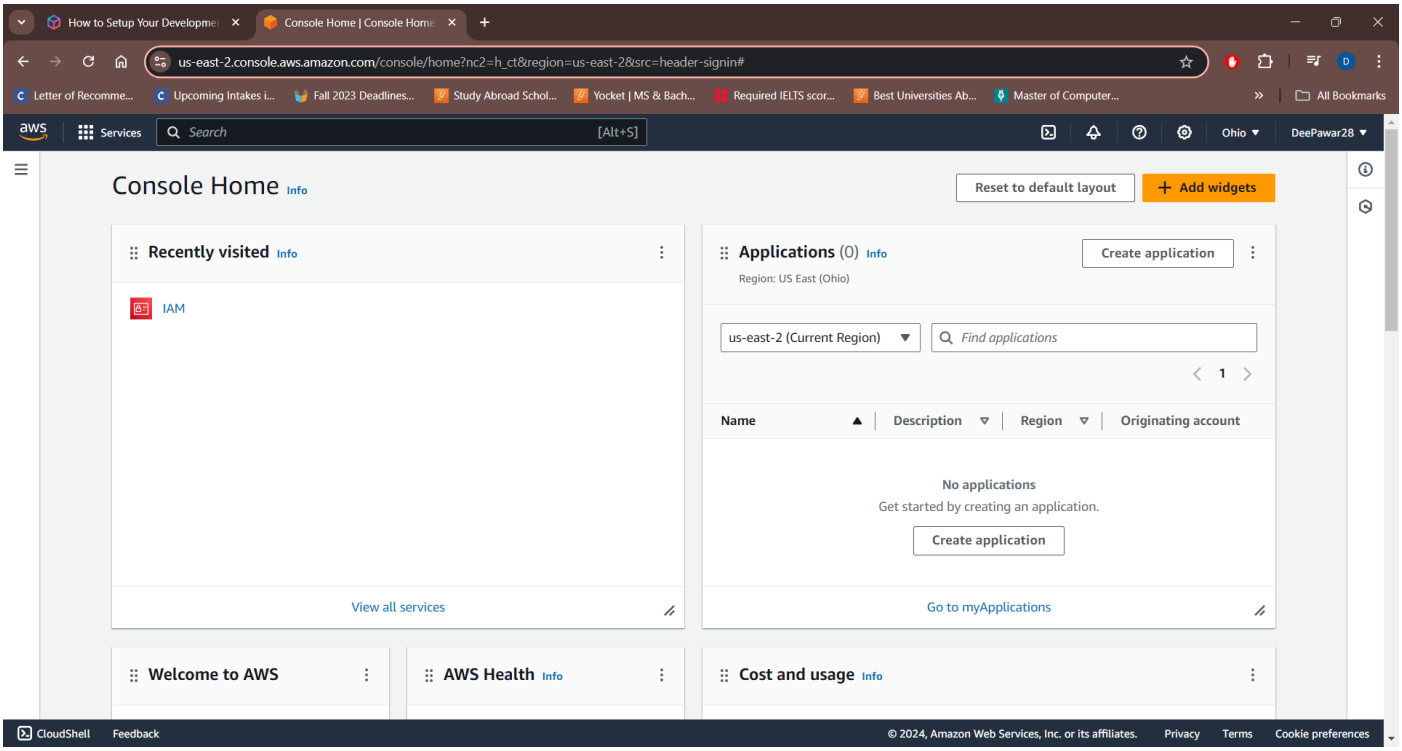
Several aspects of algorithm dynamics impacted GFT, these include regular modifications to Google's search algorithm and improvements to enhance user experience, Google's natural cultivation of search behavior, and the implementation of recommended searches. Each of these factors altered user search patterns, which in turn changed the data that GFT depended on. Although these adjustments were made as part of Google's efforts to enhance its offering and support its business model, they had unexpected effects on programs like GFT that depend on the consistency of search trends over time.

5. (3 points) Set up an Amazon Web Services (AWS) cloud account, if you don't already have one (see below for details). Since we will do most of our assignments using AWS, this will get you started.

To get credit for this part of the assignment, provide a screenshot of the main page of the AWS management console page including your account name.

Ans:

- **AWS account created:**

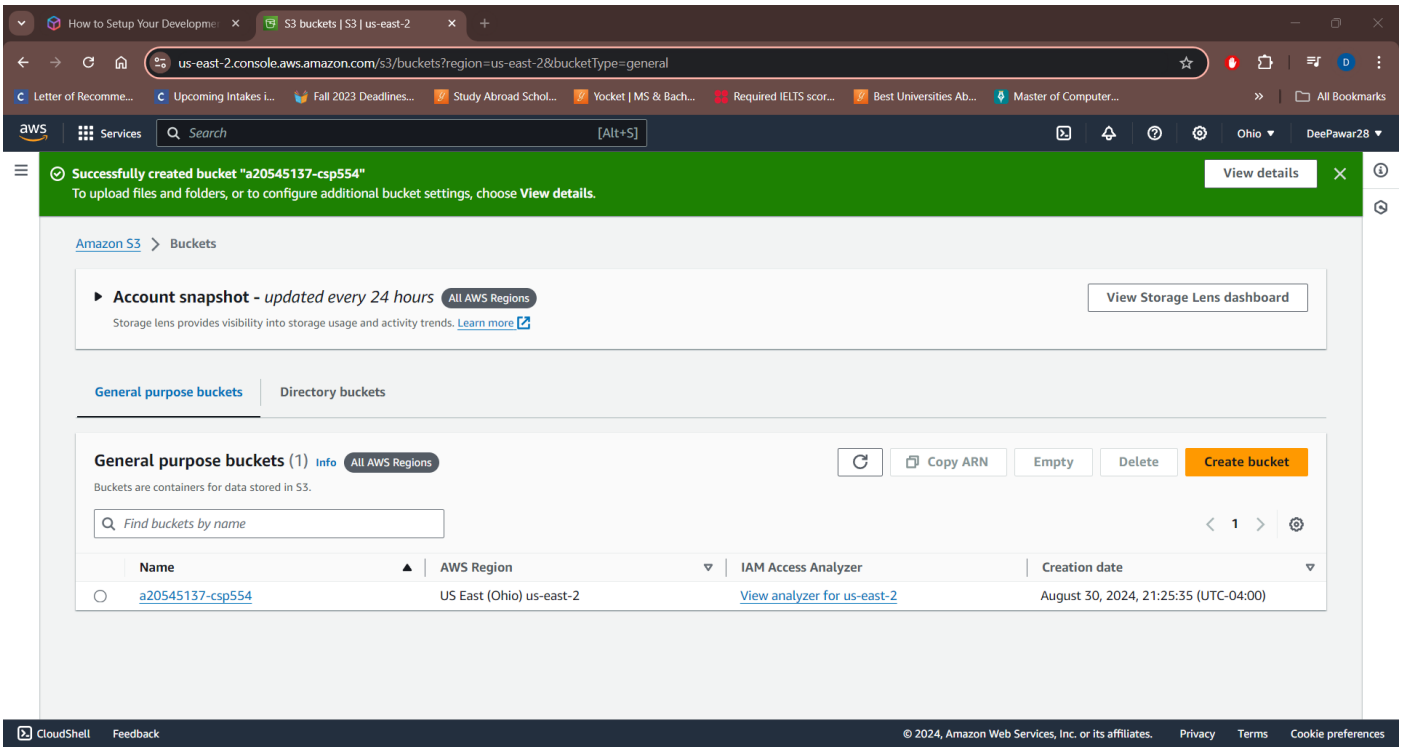


6. (2 points) Now follow the below steps about how to work with an AWS object storage service called S3 (Simple Storage Service). In a while, we will come to understand S3 as one critical element of a big data processing architecture known as the “data lake.

To receive credit for this question, provide a screenshot showing some named object is in the bucket. Note, this is the screen that appears after you choose the Upload button.

Ans:

- **S3 bucket created and named “a20545137-csp554”:**




- **Uploaded new object “CSP 554 Assignment 1 v6-1.docx” in s3 bucket:**