Kajol Tanesh Shah A20496724 kshah127@hawk.iit.edu CSP 554 Fall 2022

CSP 554 - Big Data Technologies Assignment #1

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

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

Ans. The problem with the Google Flu detection algorithm was that it overestimated the prevalence of flu i.e. it was predicting more than double the proportion of doctor visits for influenza-like illness(ILI) than the Centers for Disease Control and Prevention(CDC), which based its estimates on surveillance reports from laboratories across the United States. GFT used relative prevalence of search terms in its model and so, any changes in the search algorithm adversely affected GFT's estimates.

Below are the issues that contributed to GFT's mistakes:-

- 1) Big Data Hubris and
- 2) Algorithm Dynamics
- 2. What is big data hubris?

Ans. Big data hubris is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis. There are enormous scientific possibilities in big data, however, the quantity of data does not mean that we can ignore foundational issues of measurement and construct validity and reliability, and dependencies among data.

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

Ans. Google Flu detection algorithm could have been combined with other near real-time health data for improvement. By combining GFT and lagged CDC data, as well as dynamically recalibrating GFT, the performance of GFT could have been substantially improved. Also, Google developers could have changed the methodology of GFT by focusing on why the trending flu-related keywords were searched for instead of the number of the trending keywords.

4. What is "algorithm dynamics?"

Ans. Algorithm dynamics are the changes made by engineers to improve commercial service and by consumers in using that service.

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

Ans. Algorithm dynamics affected Google's search algorithm and this is what likely led to GFT becoming an unstable reflection of the prevalence of the flu. Several changes in Google's search

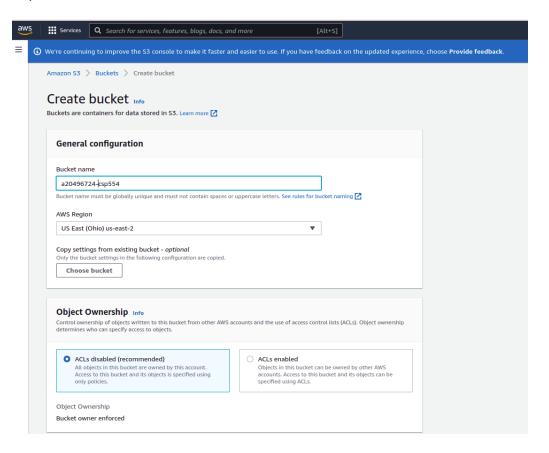
algorithm and user behavior likely affected GFT's tracking. Also, the explanation for changes in the relative search behavior was blue team dynamics where the algorithm producing the data was modified by the service provider in accordance with their business model. Because GFT used the relative prevalence of search terms in its model, improvements in the search algorithm adversely affected GFT's estimates.

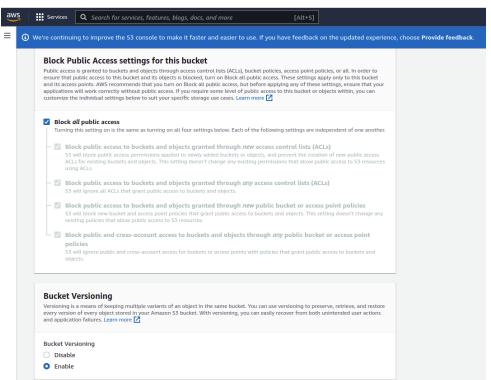
Q. (5 points) Set up an Amazon Web Services (AWS) cloud account, if you don't already have one (see below for details), and then follow the tutorial about how to work with a storage service called S3. Since we will do most of our assignments using AWS, this will get you started. In a while we will come to understand S3 as one critical element of a big data processing architecture know as the "data lake."

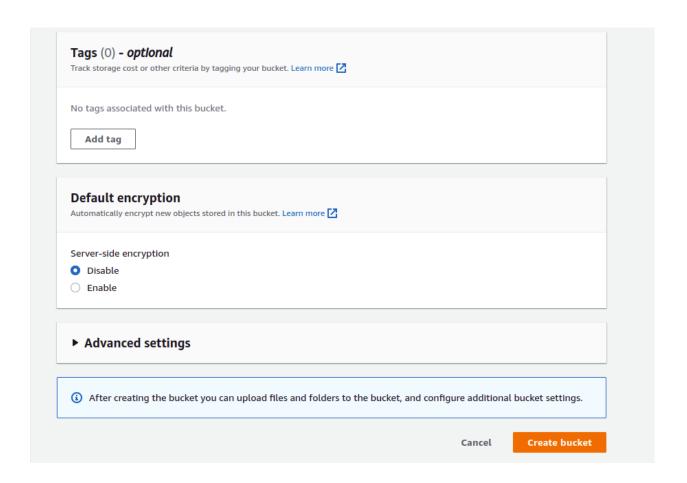
- a. To receive credit for this question, provide a screenshot showing the S3 bucket you have created.
 The bucket name should be named something like "YourIITId-CSP554", for example:
 "A1234567_CSP554"
- b. When asked to upload an object to the S3 bucket you have created, just use any text file you have handy (even this one).
- c. Now also provide a screenshot showing some named object is in the bucket.
- d. Make sure to follow the instructions in the pdf file for deleting your bucket at the end of the assignment so you do not incur additional costs.

Solution:-

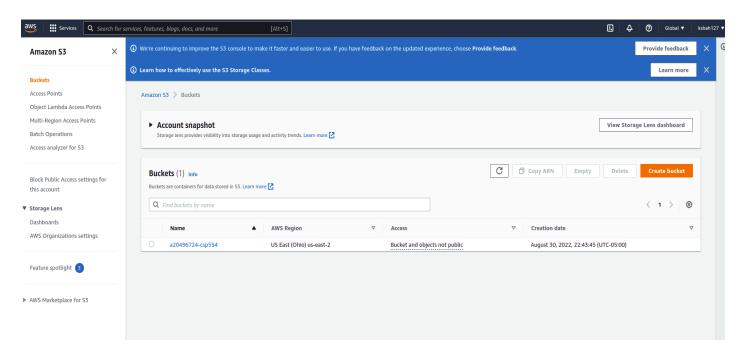
Step 1:- Create bucket



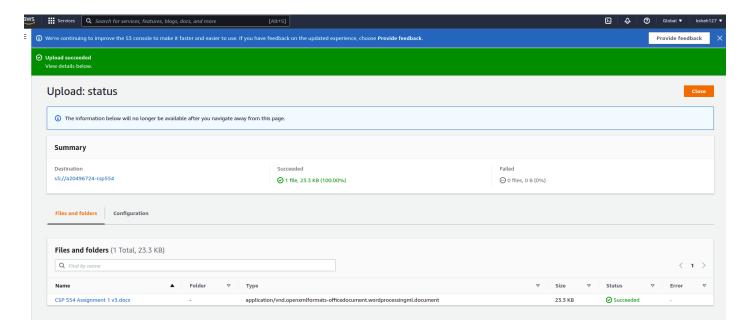




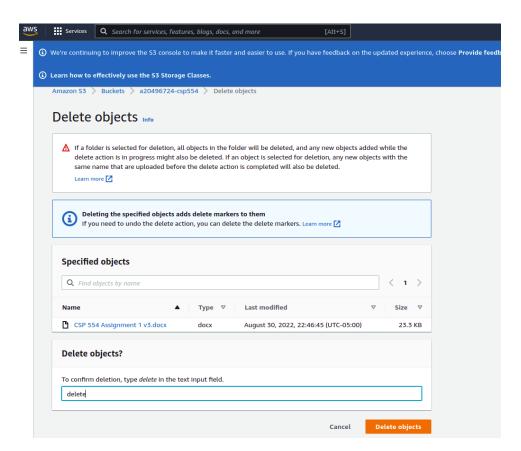
Step 2:- Bucket created

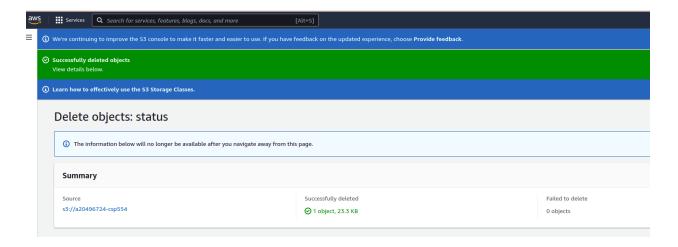


Step 3:- Upload an object to S3 bucket



Step 4:- Delete object





Step 5:- Delete bucket

