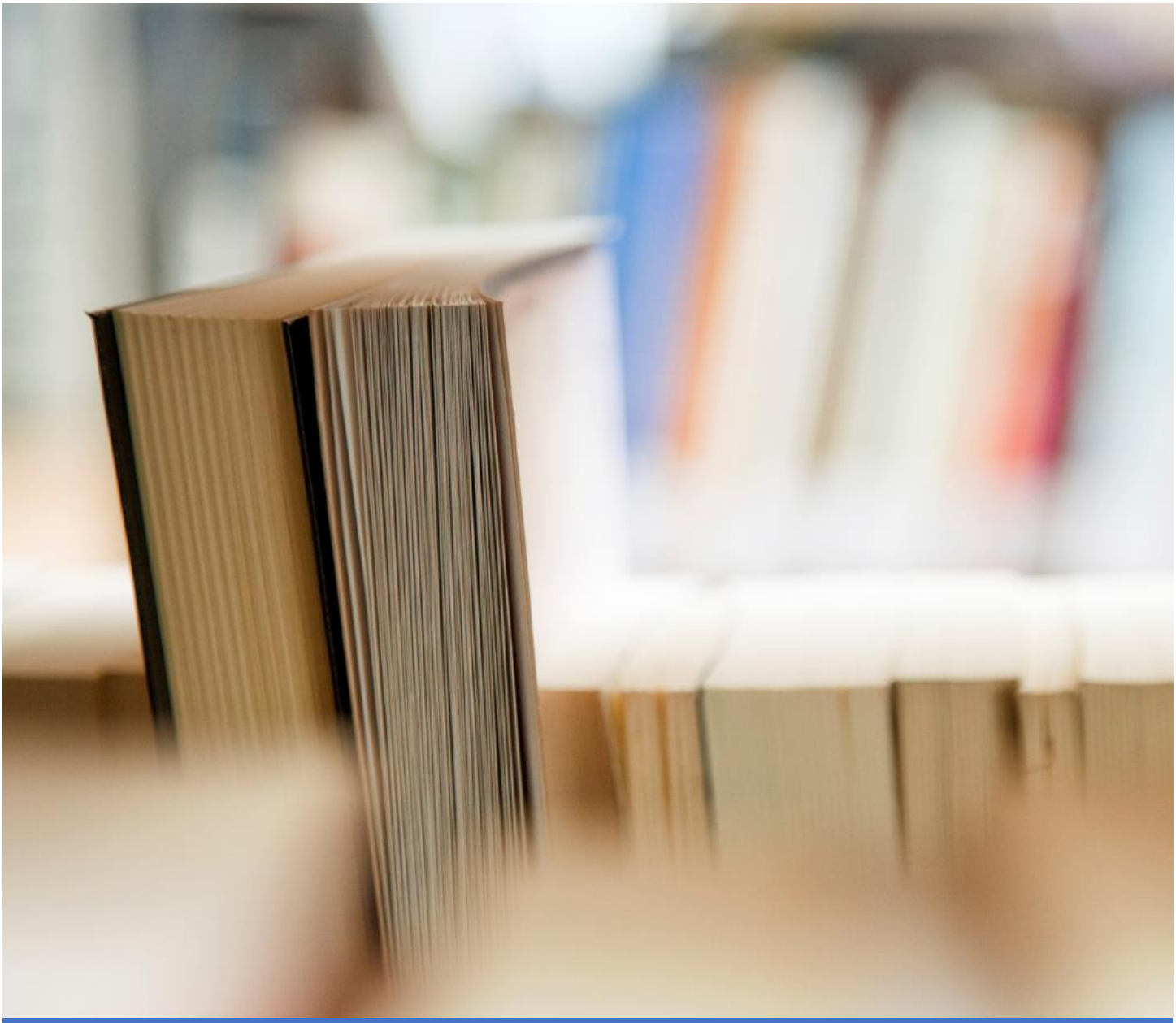


Assignment 1

Ashmita Gupta (A20512498)

CSP 554

1. Read an article on “Blackboard” in “Articles” section

The Parable of Google Flu (just 3 pages!)

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

- **What was the problem with the Google flu detection algorithm?**

Ans. GFT or Google flu detection algorithm was predicting more than double the proportion of doctor visits for influenza-like illness than the Centers for Disease Control and Prevention (CDC), which were based on its estimates on surveillance reports from laboratories across the United States. Basically, GFT was over-predicting the flu by a significant margin and its inaccuracies raised concerns about the reliability of big data analytics for public health surveillance. Two issues that contributed to GFT’s mistakes:

1) Big data hubris
2) Algorithm dynamics

- **What is big data hubris?**

Ans. Big data hubris is commonly an implicit assumption that big data is a replacement, rather than a supplement to traditional data collection and analysis. It fails to consider the limitations of big data, such as its susceptibility to bias and noise. Further, the quantity of data does not mean that one can ignore fundamental measurement issues and build validity, reliability, and dependencies between data. Therefore, to avoid big data hubris, it is important for us to be aware of the limitations of big data and to use it in combination with other methods.

- **What approach could have been used to improve the Google flu detection algorithm?**

Ans. Google Flu detection algorithms could have been improved if the developers had changed the detection method from simply counting the number of flu-related keywords to understanding the why behind those searches. For example, the search term "high school basketball" is strongly correlated with the CDC data because flu cases tend to peak during the winter, when high school basketball is in season. However, the search term "high school basketball" is not actually related to the flu. It is simply a seasonal term that is correlated with the flu by chance. This should have been a warning that the big data were overfitting the small number of cases. Further, the concept of algorithm dynamics must have been carefully considered, as it is one of the main factors that can lead to an algorithm displaying erroneous data, which can result in a high percentage of error.

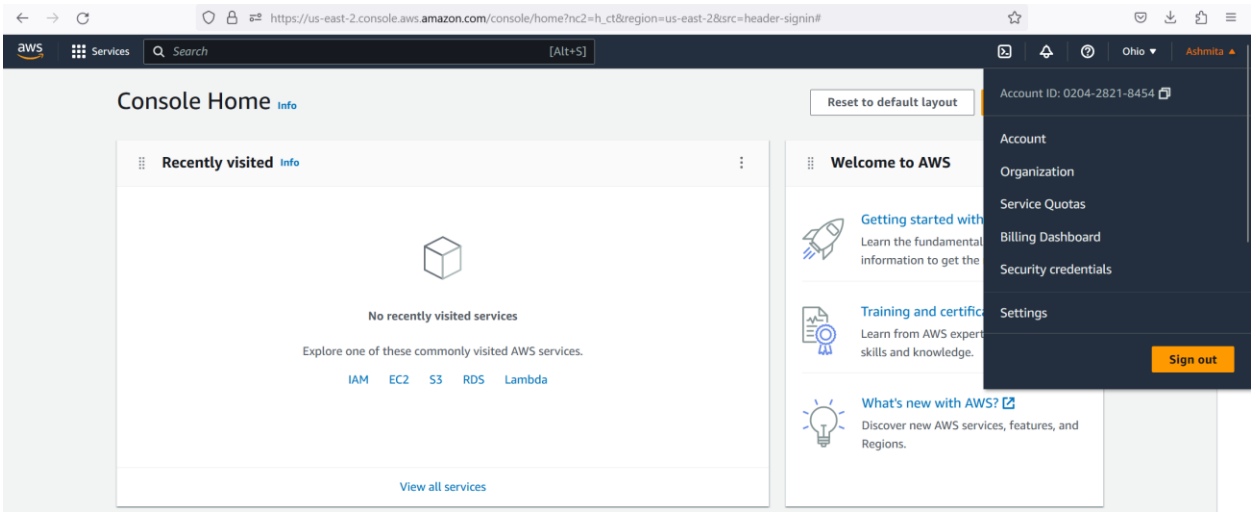
- **What is “algorithm dynamics?”**

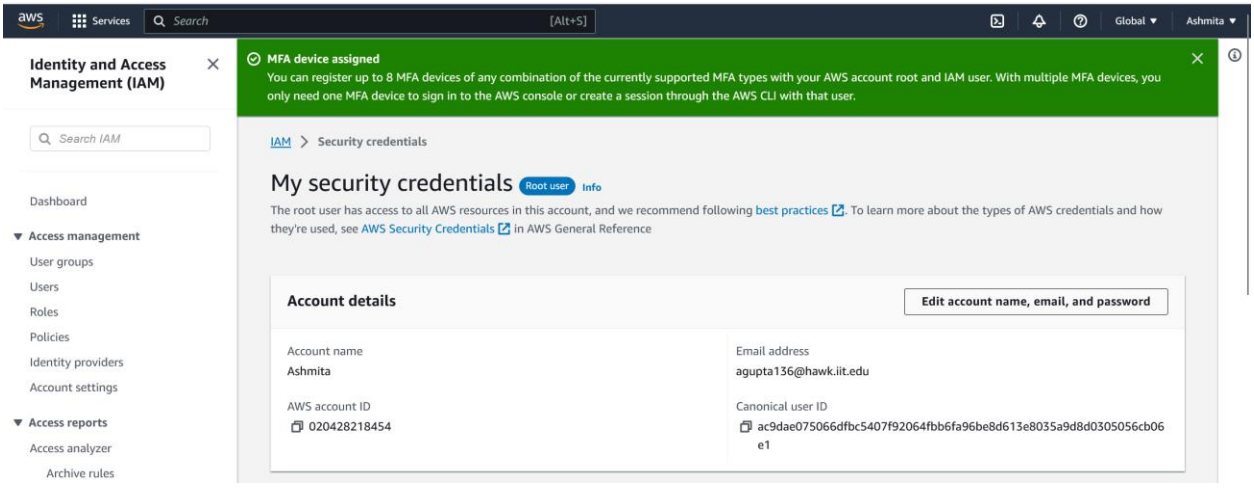
Ans. Algorithm dynamics refers to the way that an algorithm changes over time in response to new data. It is basically the changes made by engineers to improve the commercial service and by consumers in using that service. Algorithm dynamism is important because it helps the algorithm adapt to new changes based on the changes occurring over a certain period in society

- **What aspect of algorithm dynamics impacted the Google flu detection algorithm?**

Ans. The algorithm was trained on data from previous years, but search behavior can change over time. For instance, people are more likely to search for flu-related terms during a media-stoked panic. This led to the algorithm overestimating the number of flu cases. GFT uses the relative popularity of the search terms in its model and thus impacted the Google flu detection algorithm. Further, it acts abnormally by showing unstable reflection about the prevalence of the flu.

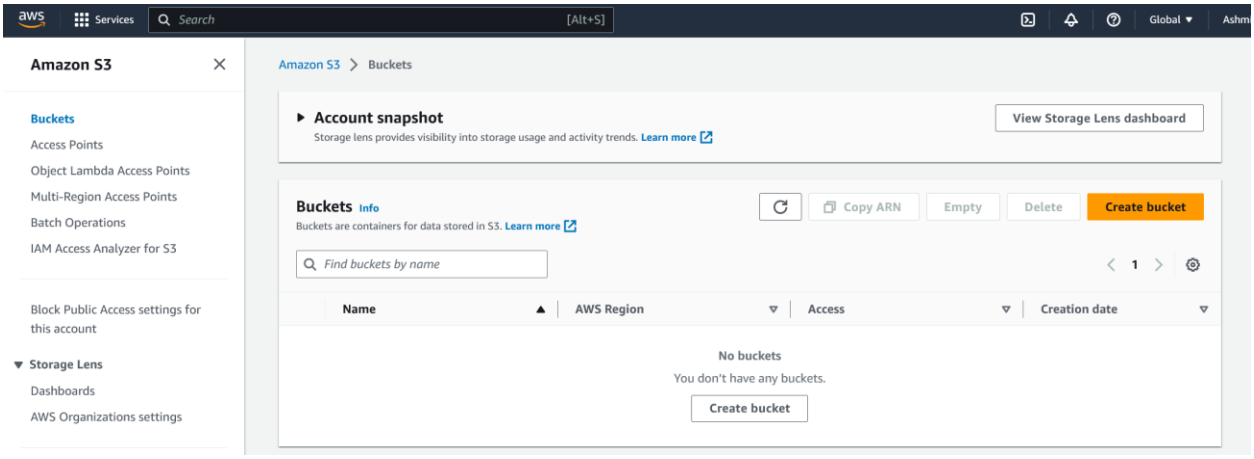
2. Amazon account creation:



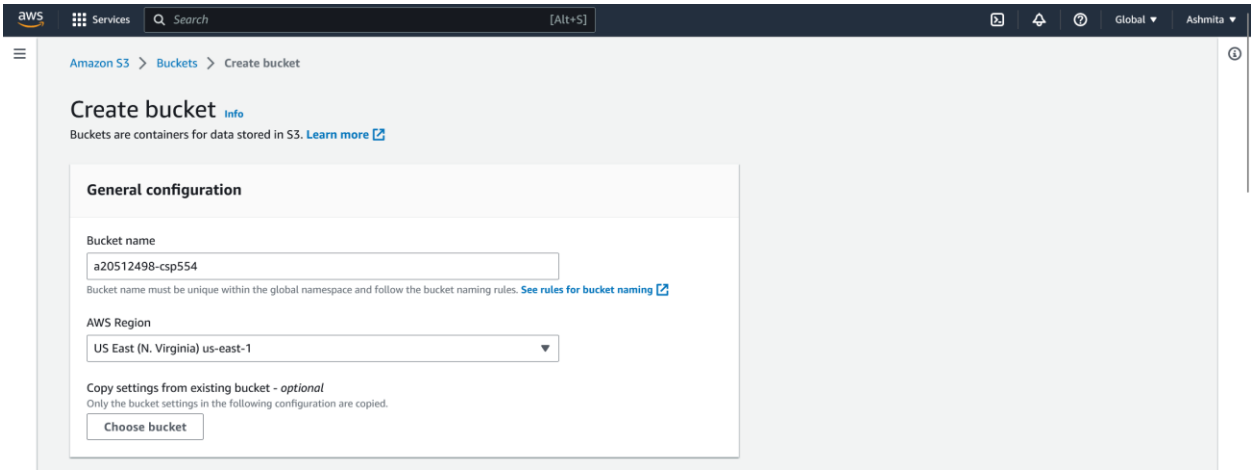


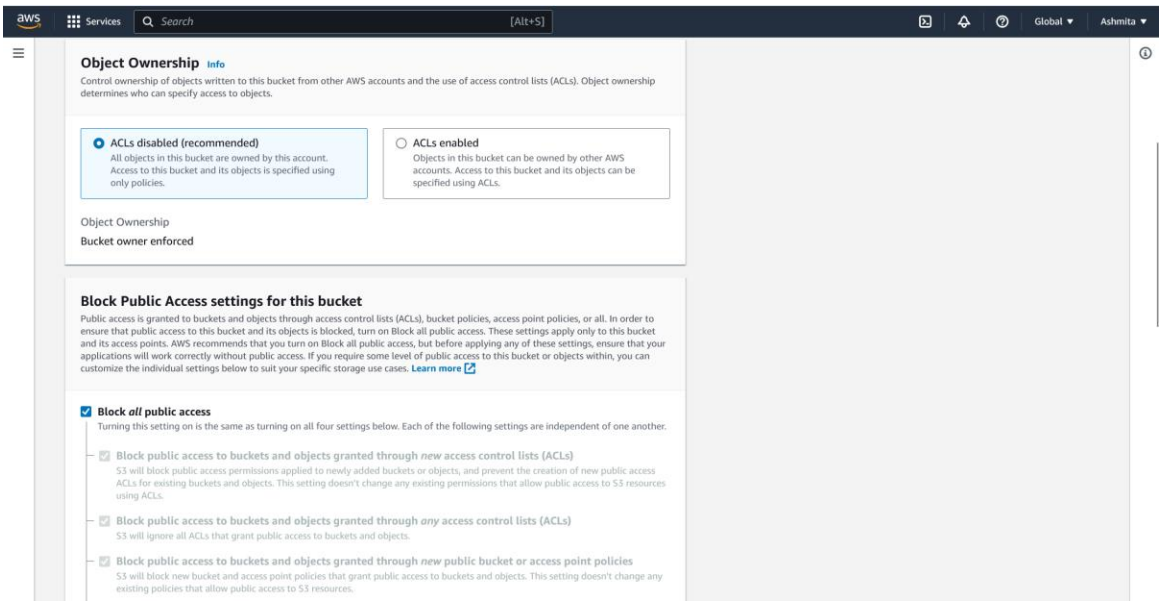
(a) Amazon S3 Bucket

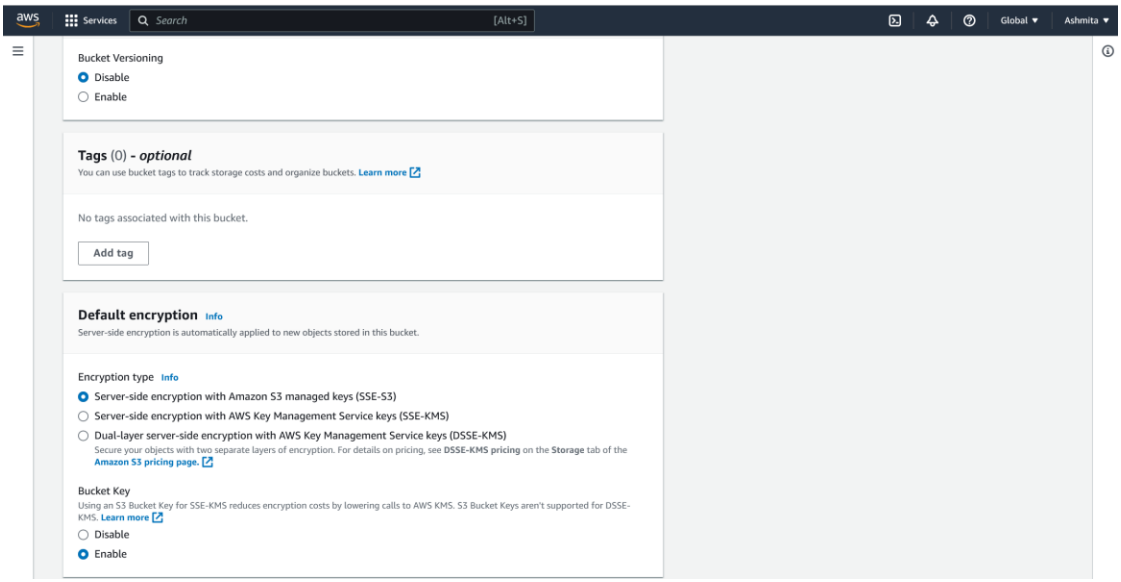
Below is the screenshot of the Amazon S3:

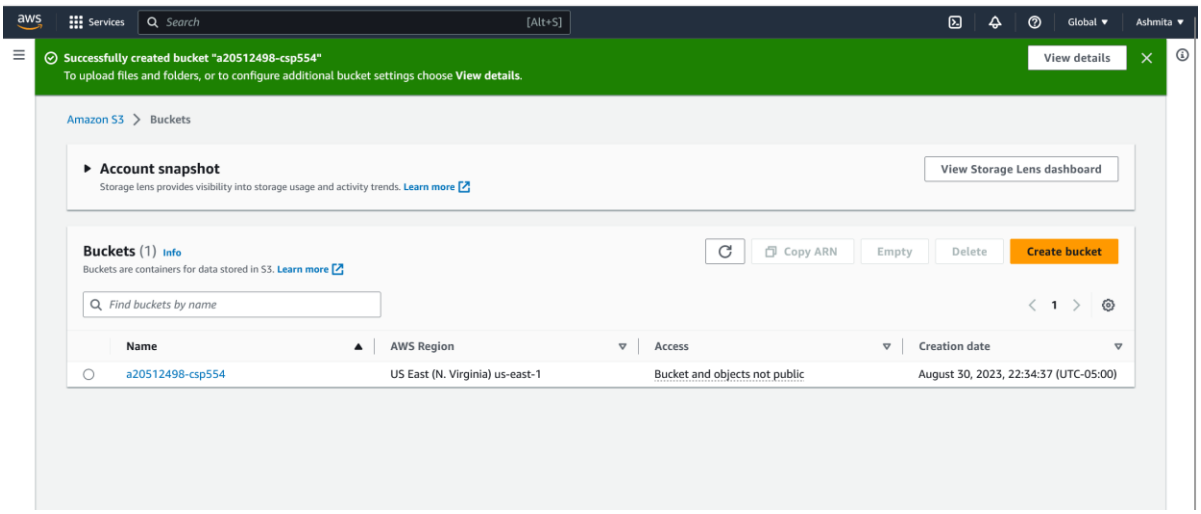


(b) Creating an S3 Bucket with the name a20512498-csp554:



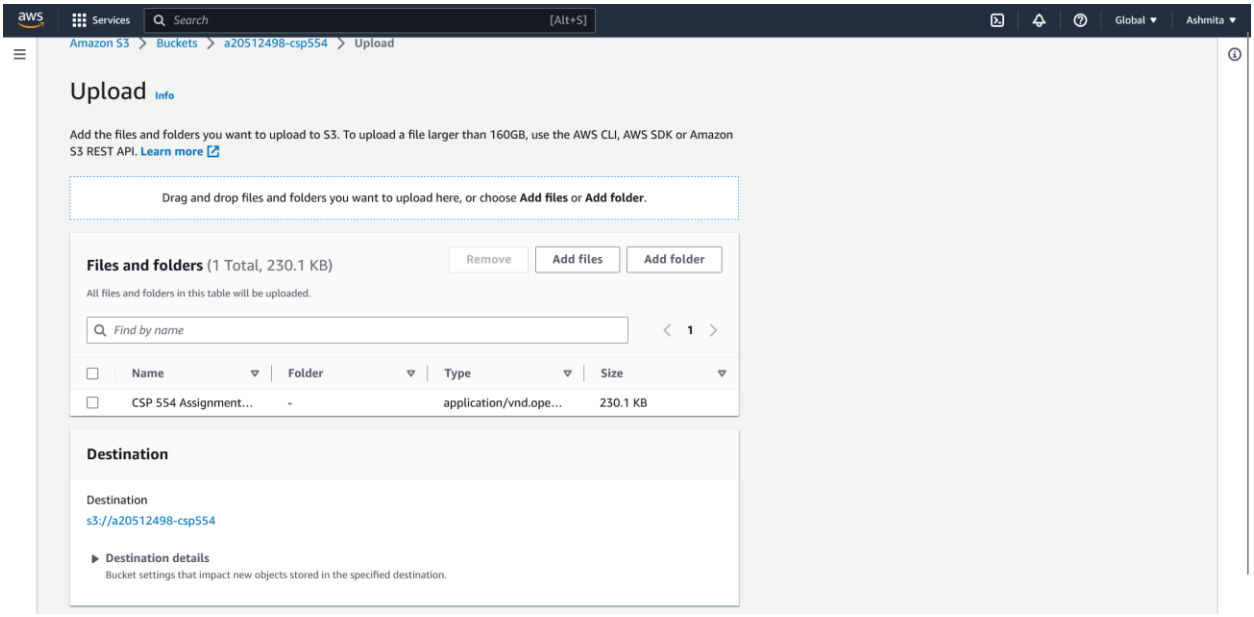


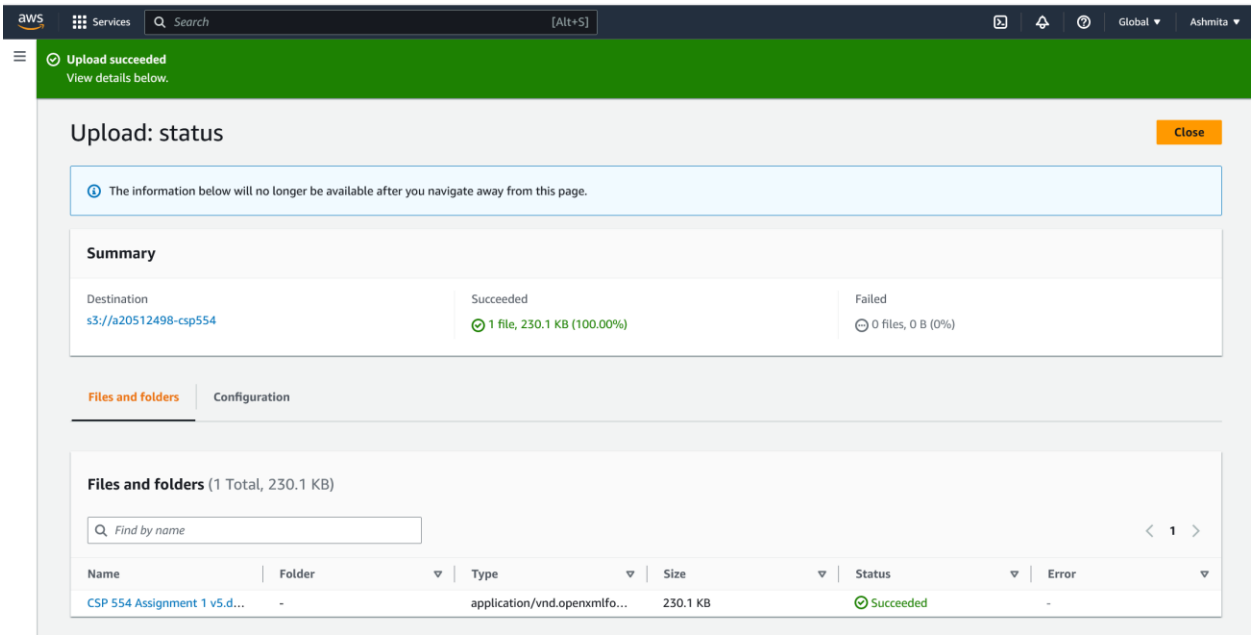




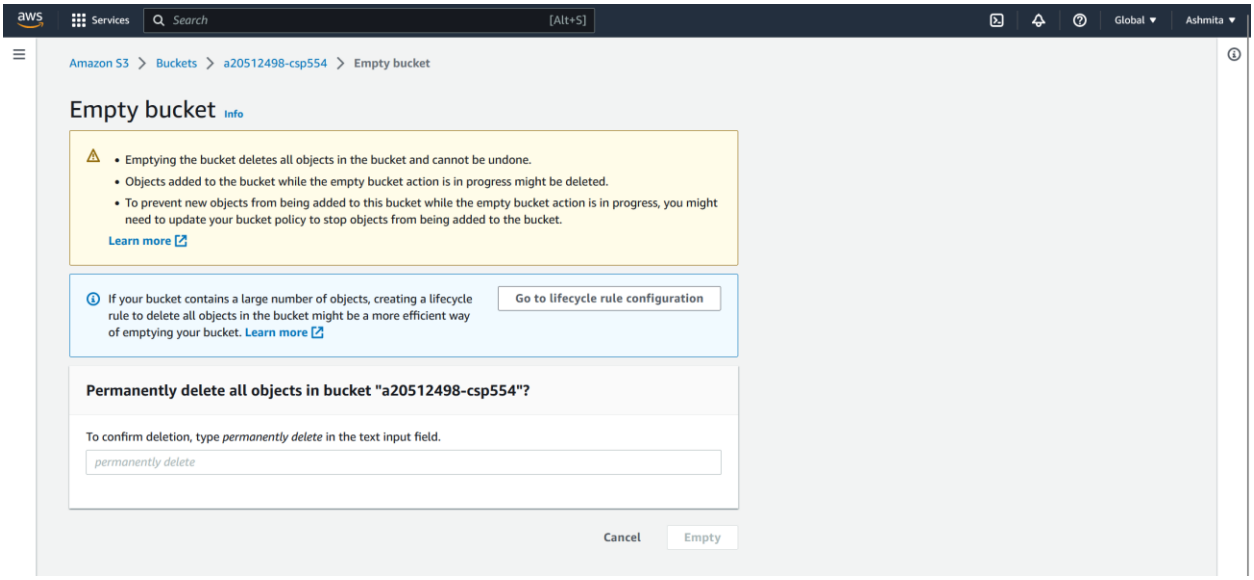
Ashmita Gupta

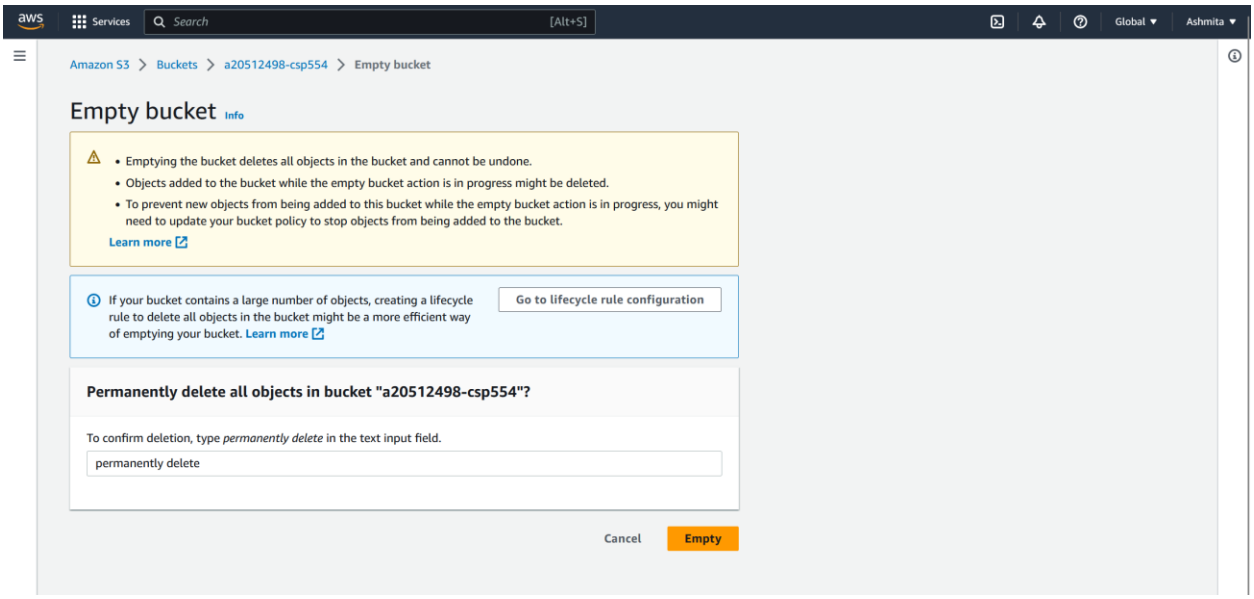
(c) Creating and uploading an object in the Bucket:

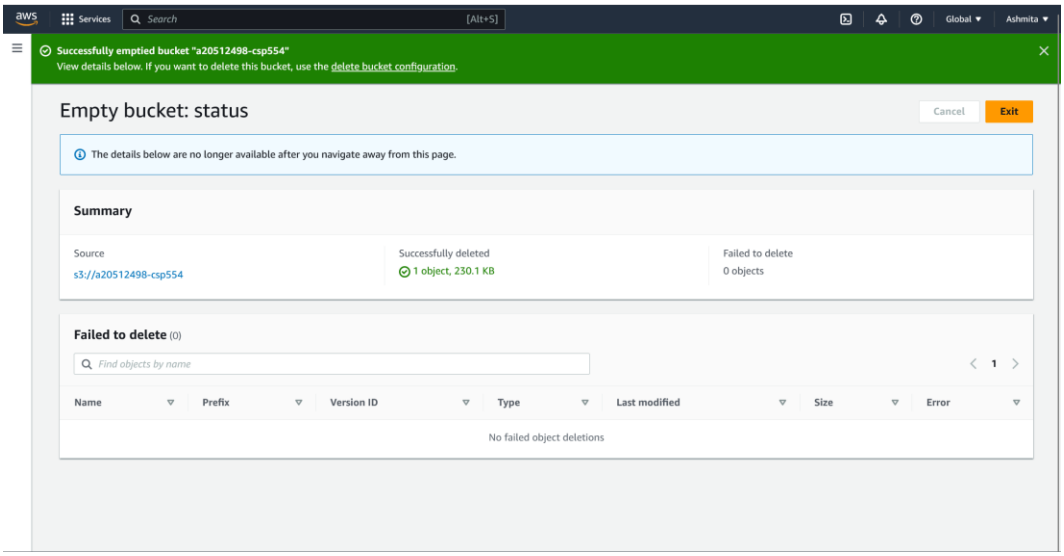


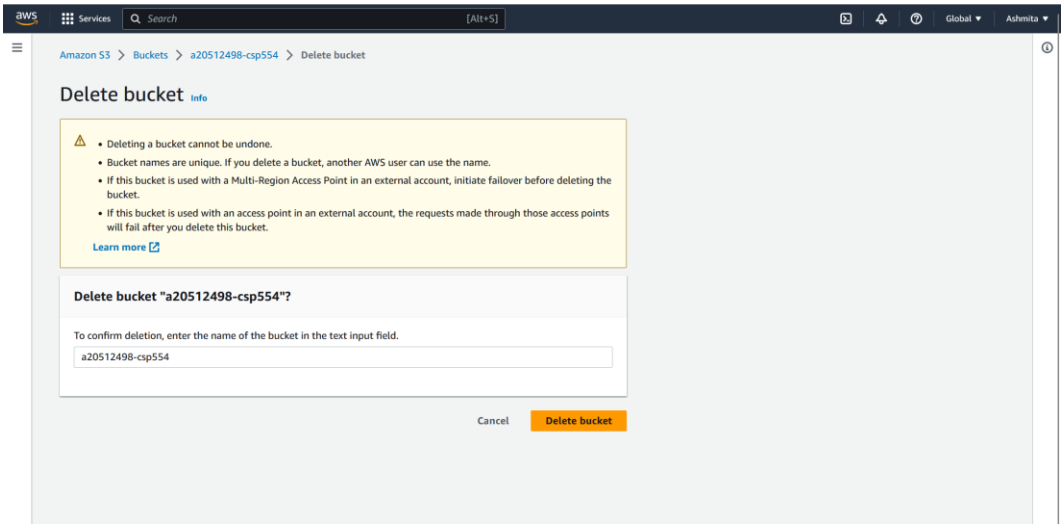


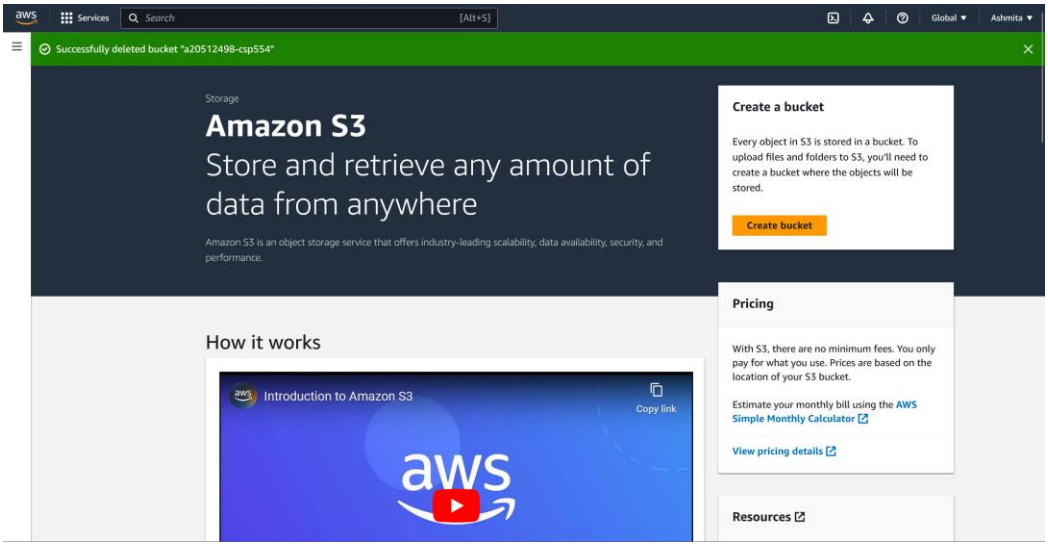
(d) Deleting a bucket and its objects:











Ashmita Gupta