CSP 554—Big Data Technologies

Assignment 1

Student ID: A20549927

4. Questions from the article

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

➔ The algorithm predicted more than double the doctor visits for flu than the Centers for Disease Control and Prevention (CDC) data.
➔ It failed despite being designed to predict CDC reports.
➔ Google Flu Trends (GFT) overfit a small number of cases, missing nonseasonal flu pandemics.

2. What is big data hubris?

➔ It's the assumption that big data can replace, rather than supplement, traditional data collection and analysis.
➔ Big data hubris overlooks foundational issues like measurement, validity, and data dependency.
➔ This hubris led to an overreliance on GFT's data without sufficient traditional methodological checks.

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

➔ Could have used traditional statistical methods to extract more information from the data.
➔ Incorporating lagged CDC data with GFT for better prediction.
➔ Continuously updating and recalibrating the algorithm based on new data and trends.

4. What is "algorithm dynamics?"

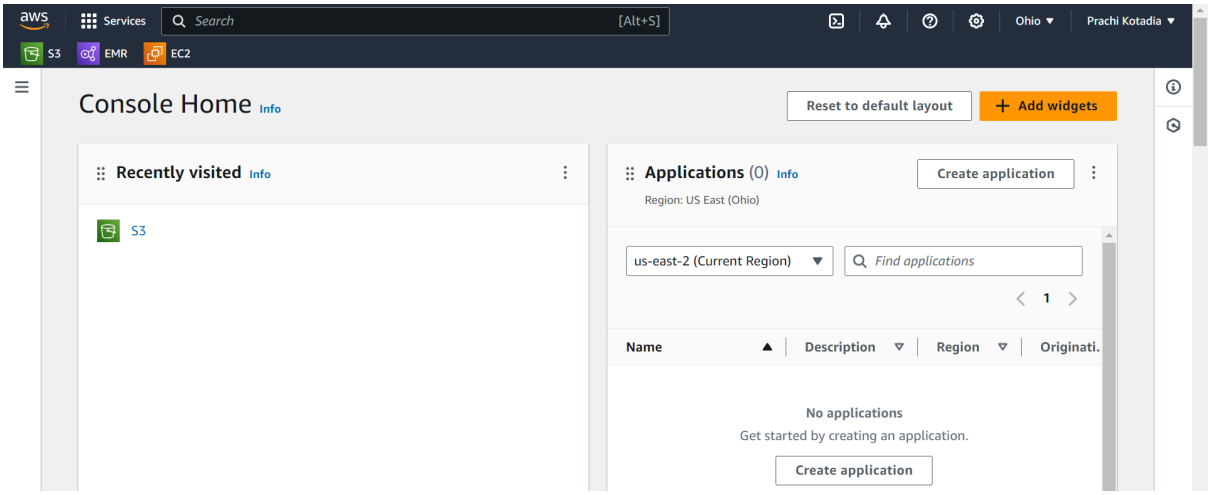
➔ Refers to changes made to improve the service and user behavior changes.
➔ These changes affect the data generation process and, consequently, the algorithm's performance.
➔ Includes both the modifications by engineers and the evolving nature of user interactions with the service.

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

➔ Changes in Google's search algorithm affected GFT's tracking accuracy.
➔ The algorithm's reliance on search terms was disrupted by the evolving search patterns.
➔ Media-driven search behavior changes, like panic during flu seasons, significantly skewed GFT's results.

5. AWS Account Screenshots:

AWS Management Console Page:



6. AWS Simple Storage Service (S3):

Screenshot of the bucket created:
A bucket named a20549927csp554 is created in the AWS console.