Chirag Khandhar

A20438926 | CSP 554 - Big Data Technologies | Fall 2020 | Assignment 1

Q3. Submit very brief answers to the following questions:

i. What location or time zone are you in when you attend the course?

- Chicago, IL (Central Daylight Time GMT-5)

ii. Describe any prior experience you might have with use of public cloud, data mining, machine learning, statistics, data science and big data.

- No prior hands-on experience with AWS S3 or EC2.
- Experience consuming certain API's with Google Cloud Platform
- Completed CS 422 Data Mining and CSP 571 Data Preparation and Analysis courses
- Experience working with Map Reduce

iii. Share any big data interests and personal learning goals for the course.

- I expect to learn and understand how big data can be managed and what all underlying technologies are used to derive some useful application out if it.
- Being a Full Stack Developer, I want to learn this data driven component as it is very crucial in developing e-commerce applications.

iv. Indicate if there are additional topics in the scope of the course of special interest to you.

- I would love to see specific techniques or technologies on connecting these Big Data Technologies with the Web Technologies, so that the end product can communicate with these BDT's.

v. Do you have any anticipated personal issues such as expected absences or other necessary accommodations with course impact? (Of course, these will be held in strictest confidence.)

- None

Q4. Read Article on "**The Parable of Google Flu**"

Q5. 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?

- One of the contributing factor in GFT's mistake is "Big Data Hubris"
- The mistake of many big data projects, the researchers noted, is that they are not based on technology designed to produce valid and reliable data amenable for scientific analysis. The data comes from sources such as smartphones, search results and social networks rather than carefully vetted participants and scientific instruments.
- Google's algorithm was quite vulnerable to overfitting to seasonal terms unrelated to the flu, but strongly correlated to the CDC data, like "high school basketball."
- With millions of search terms being fit to the CDC's data, there were bound to be searches that were strongly correlated by pure chance, and these terms were unlikely to be driven by actual flu cases or predicting future trends.
- Google also did not consider changes in search behaviour over time.
- The other factor contributing in GFT's mistake is "Algorithm Dynamics"

ii. What is big data hubris?

- It is the assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis.

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

- GFT could have been better if it was combined with other near-real-time health data.
- For example, combining GFT and lagged CDC data, as well as dynamically recalibrating GFT.
- By doing so, GFT could have significantly improved.

iv. What is "algorithm dynamics?"

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

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

- Continuous changes are made by the Google in the search algorithm to enhance the customer satisfaction henceforth affecting the tracking capacity of GFT and resulting as a drawback to Google Flu.

Q6. 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 known as the "data lake."

- To receive credit for this question, provide a screen shot showing the S3 bucket you have created. The bucket name should be named something like "YourIITId-CSP554"
- When asked to upload an object to the S3 bucket you have created, just use any text file you have handy (even this one).
- Now also provide a screen shot showing some named object is in the bucket.

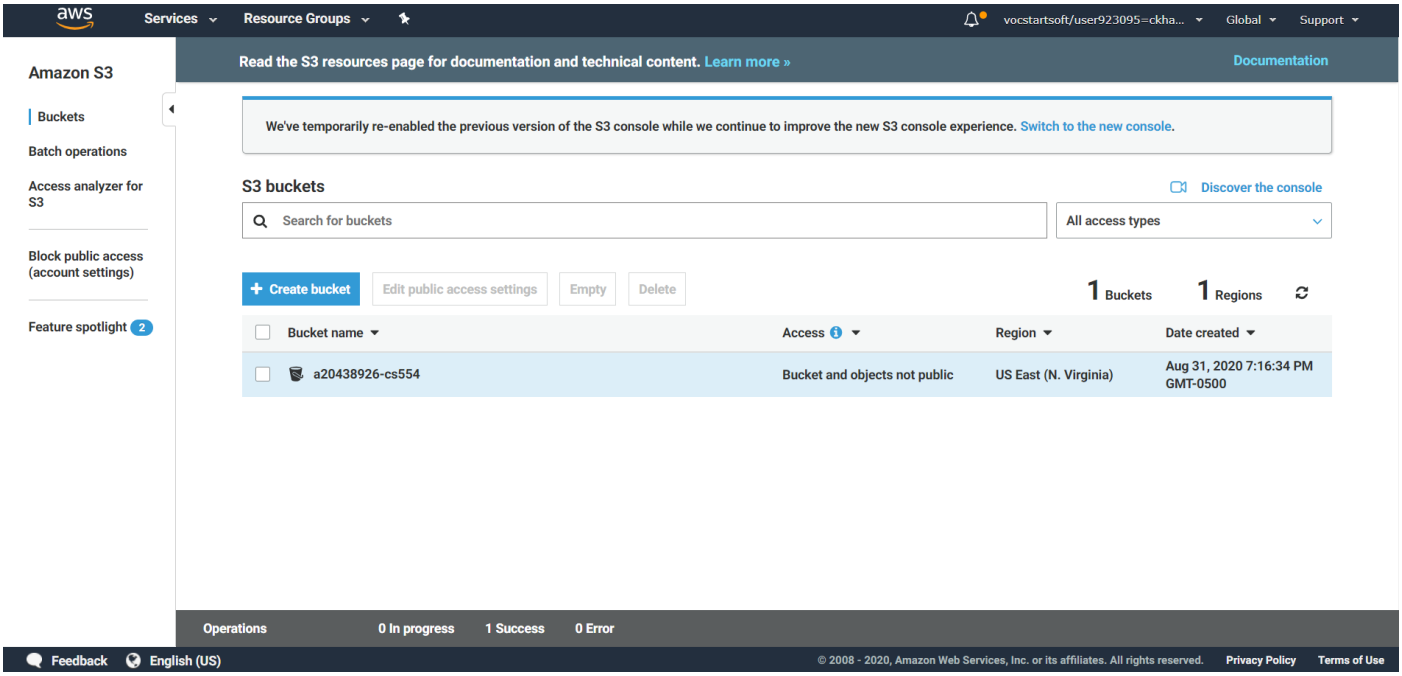

Figure 1. Bucket Created (a20438926-cs554)

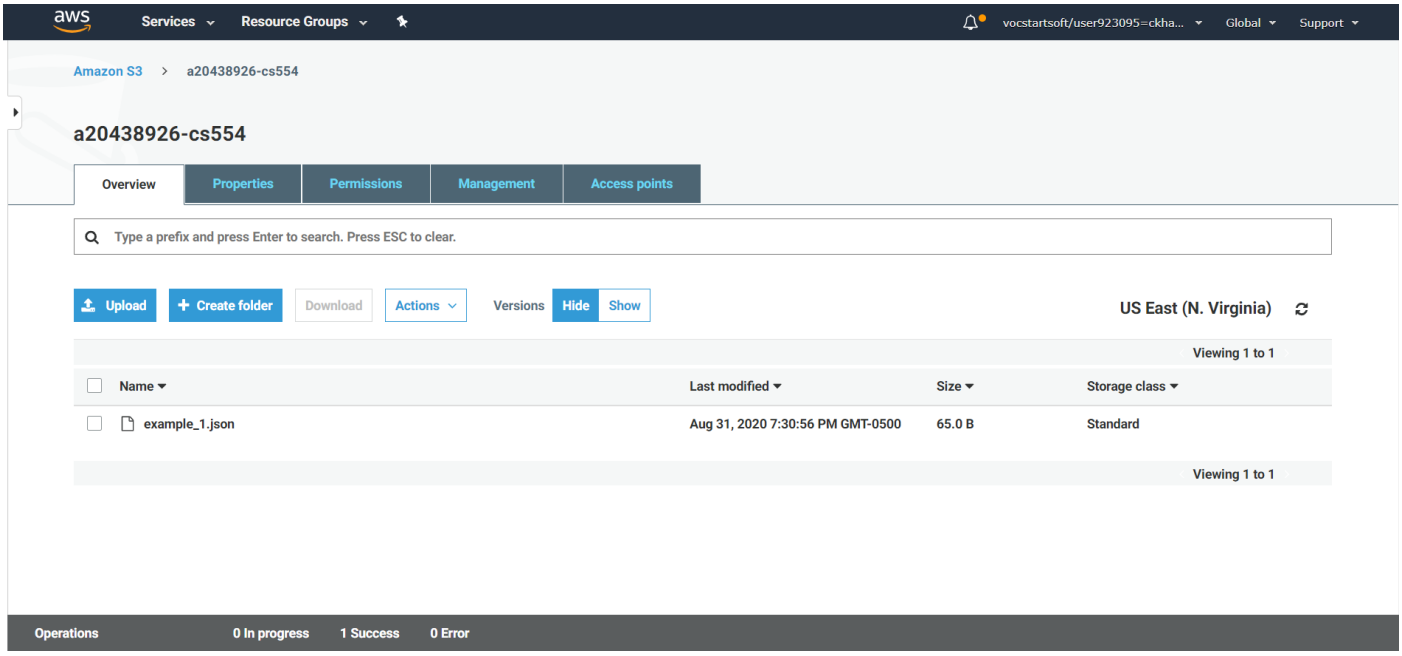

Figure 2 Object Added (example_1.json)