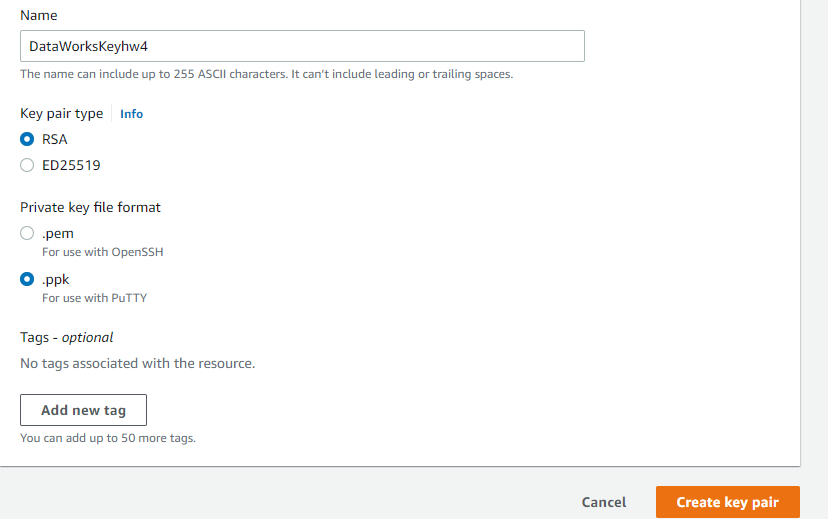
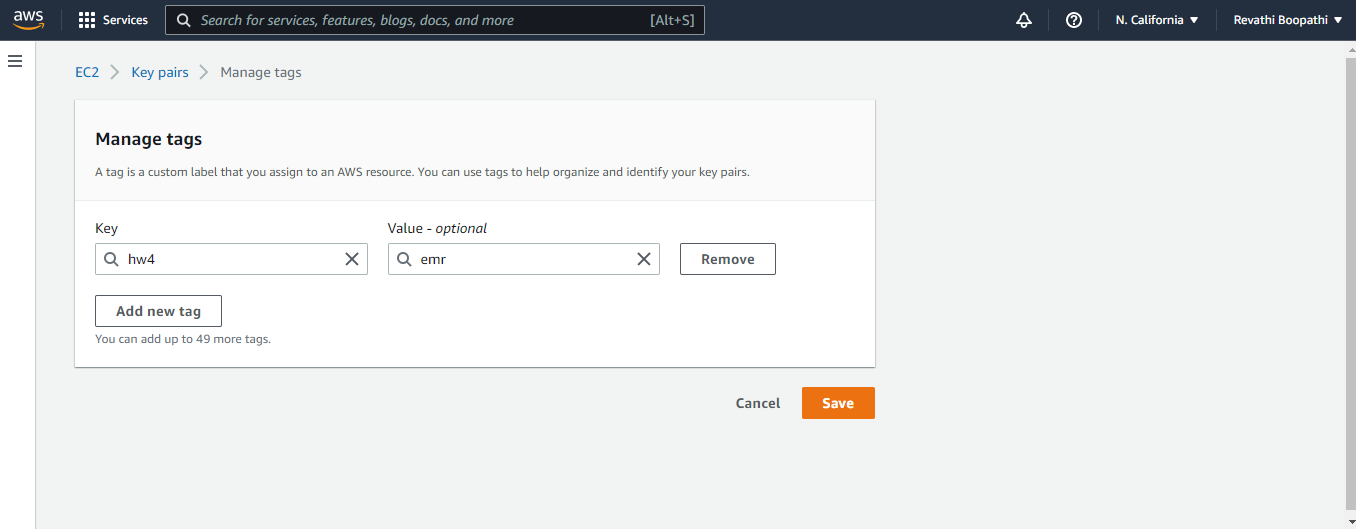
**Step 1: Set-up prerequisites**

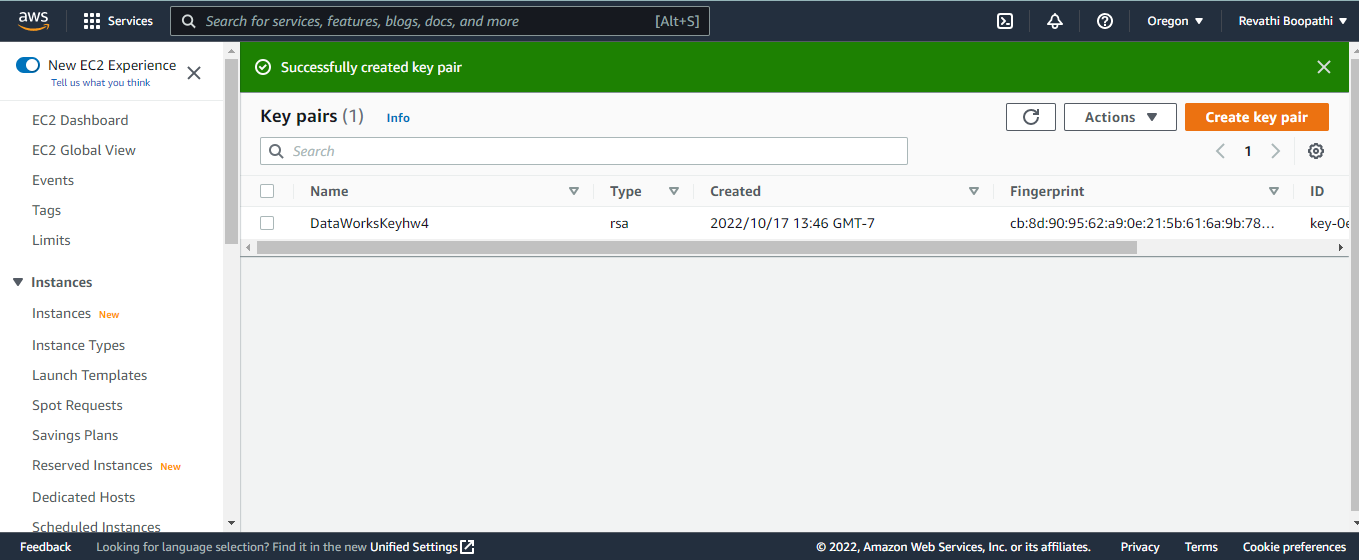
* Create an Amazon EC2 key pair for SSH, Use Amazon EC2 to create key pair



* Tag a public key with custom metadata

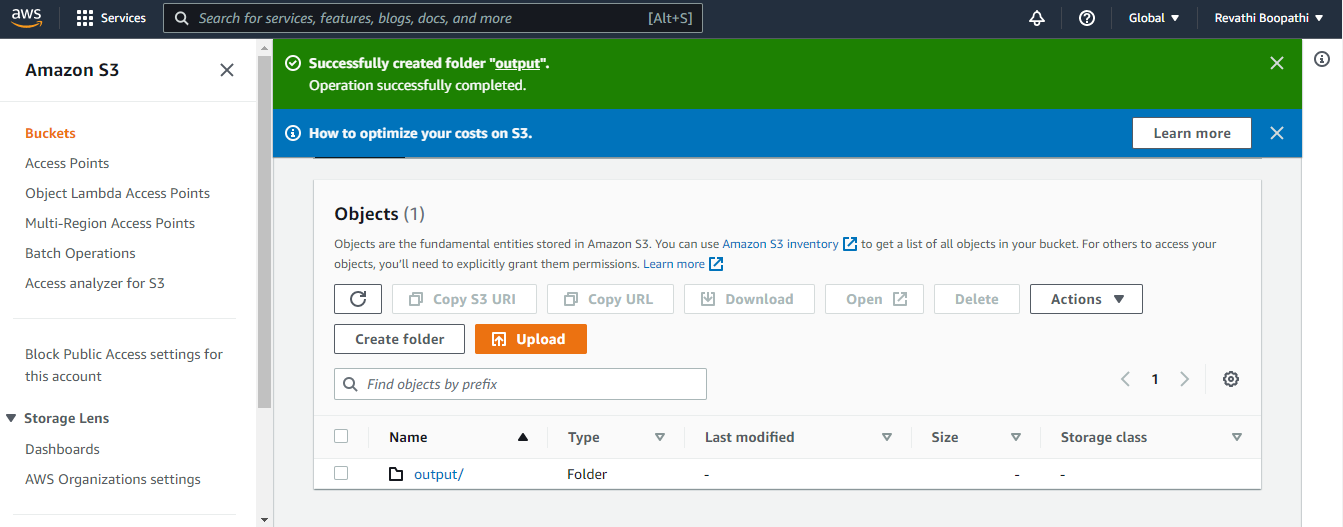


# Check the description of public key created

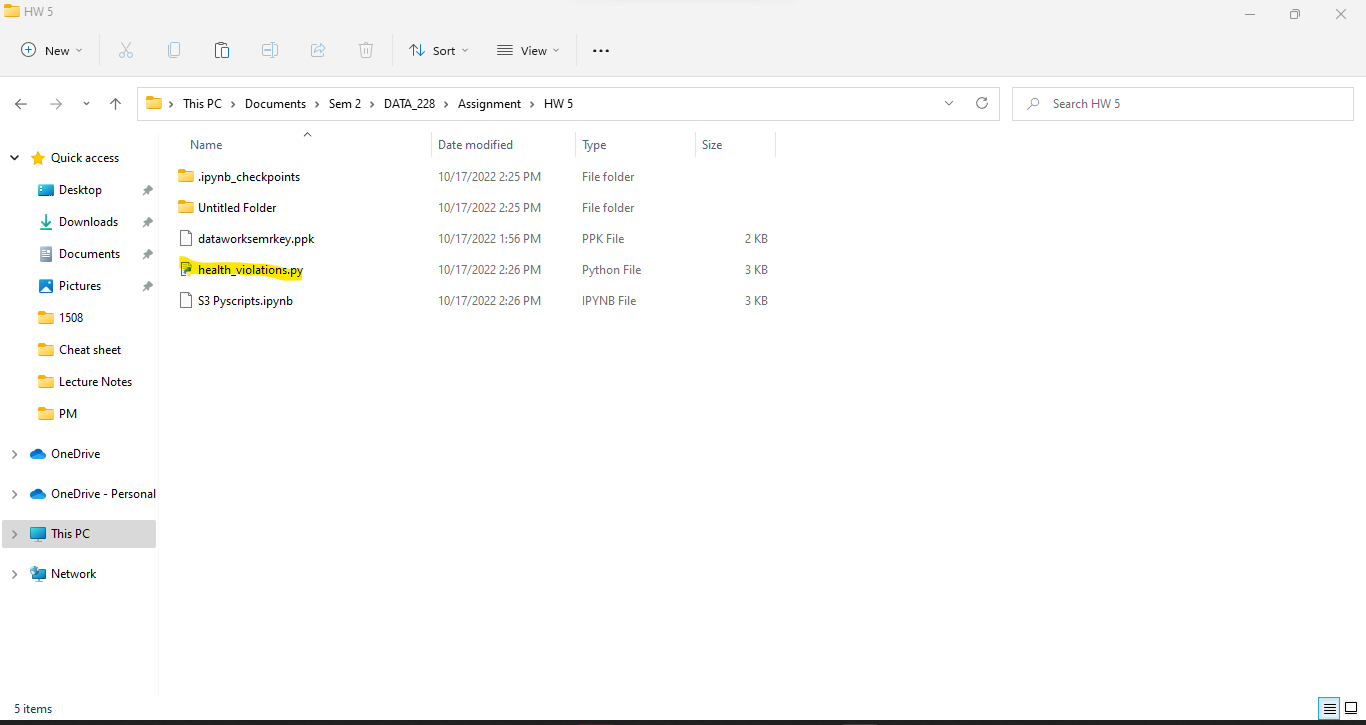


**Step 2: Launch the cluster**

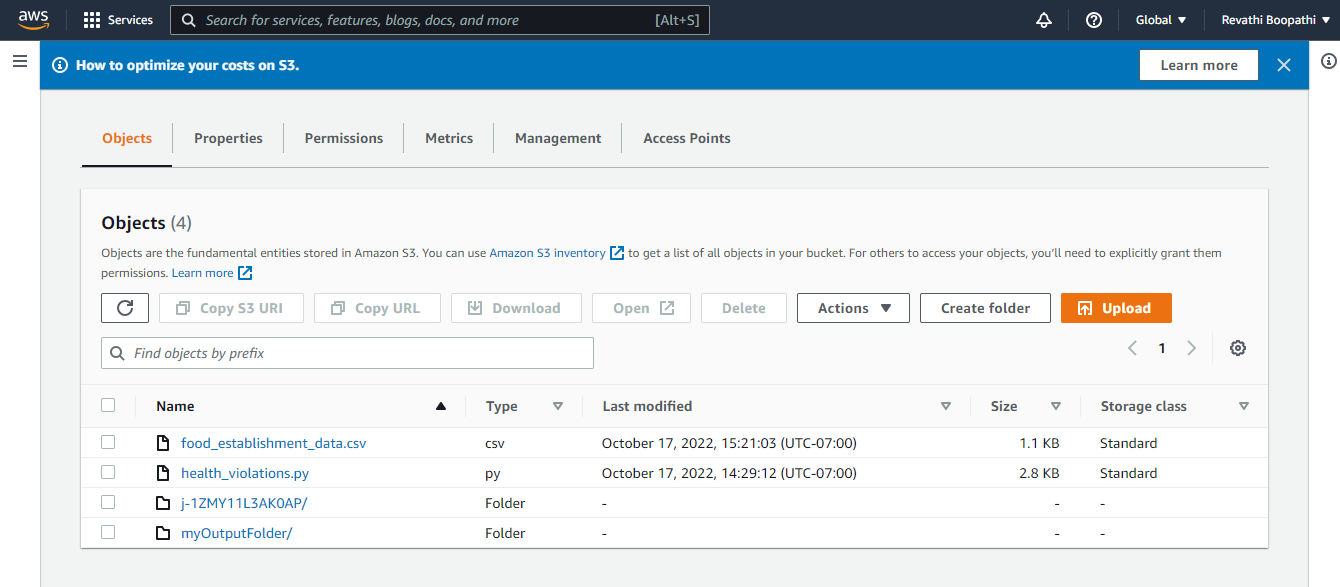
* Create an AWS S3 bucket which will store the data that would be queried
* Make sure to create an empty folder where the results would be stores



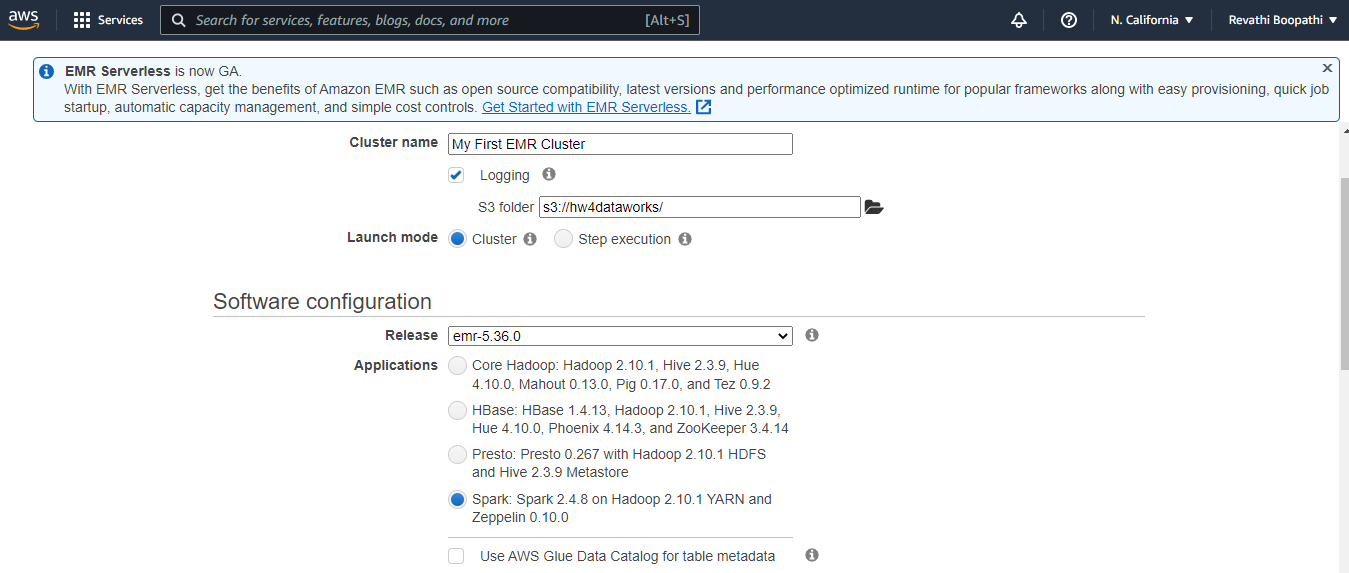
* Prepare the example PySpark script for EMR, which will query data from the sample health food samples data, paste the sample data provided in AWS documentation and paste it in a preferred editor (PyCharm), save it in .py format

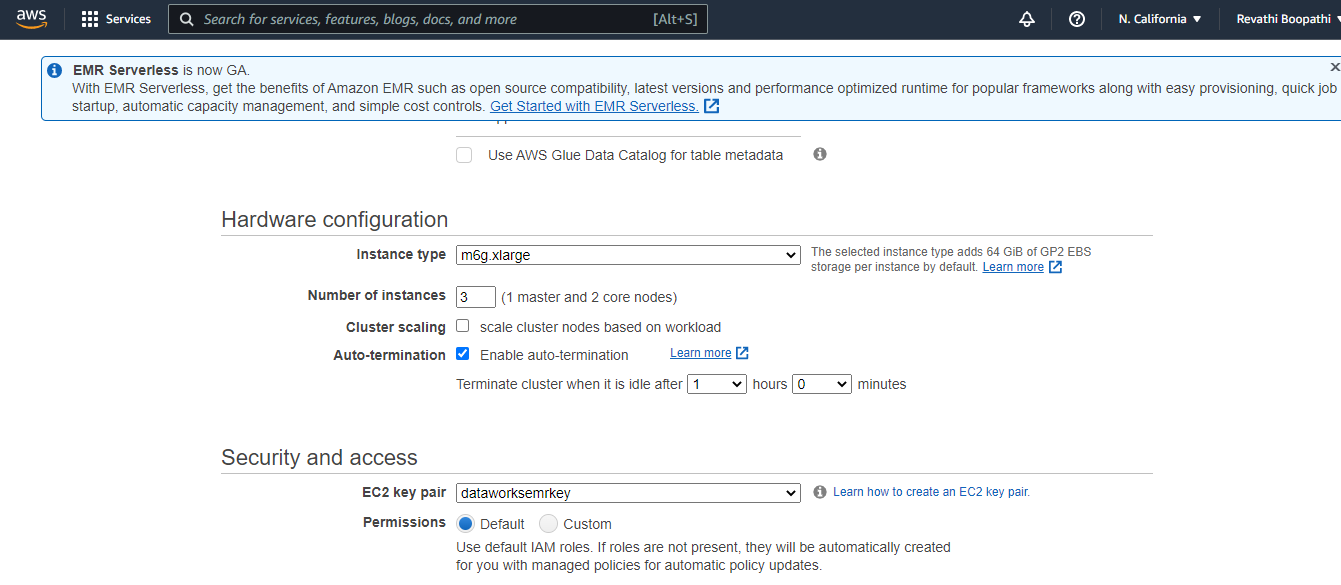


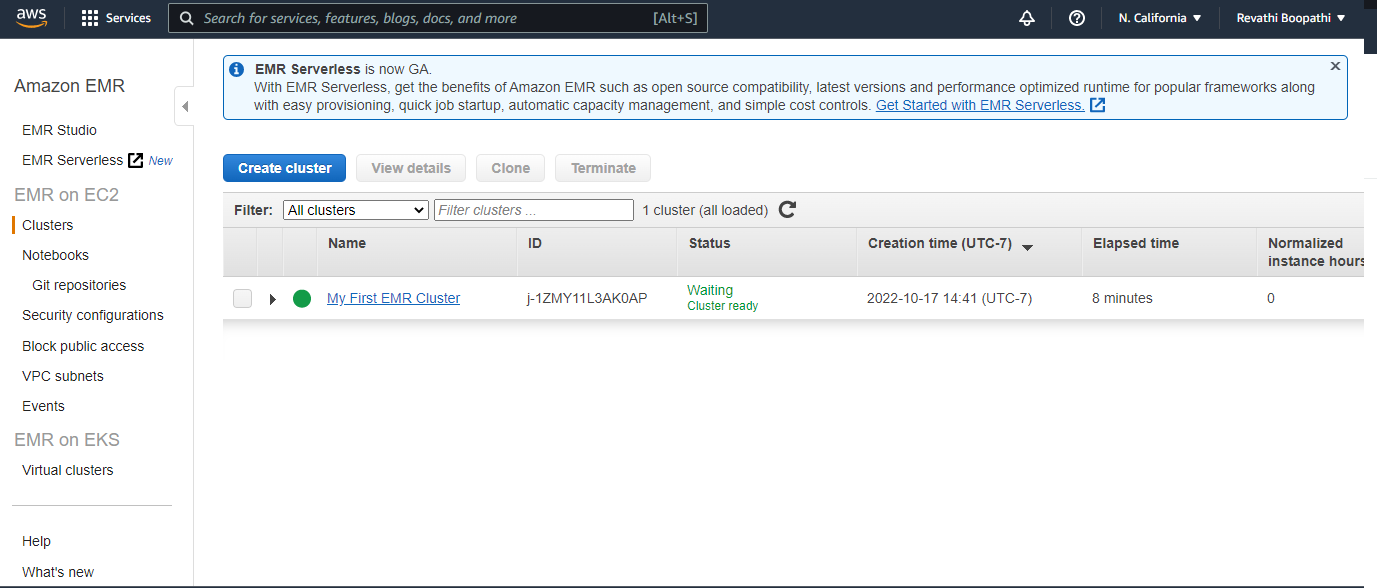
* Upload the .py file to the created S3 bucket
* Upload the sample food establishment data also to the S3 bucket



### Now, create the Amazon EMR cluster by selecting the below fields

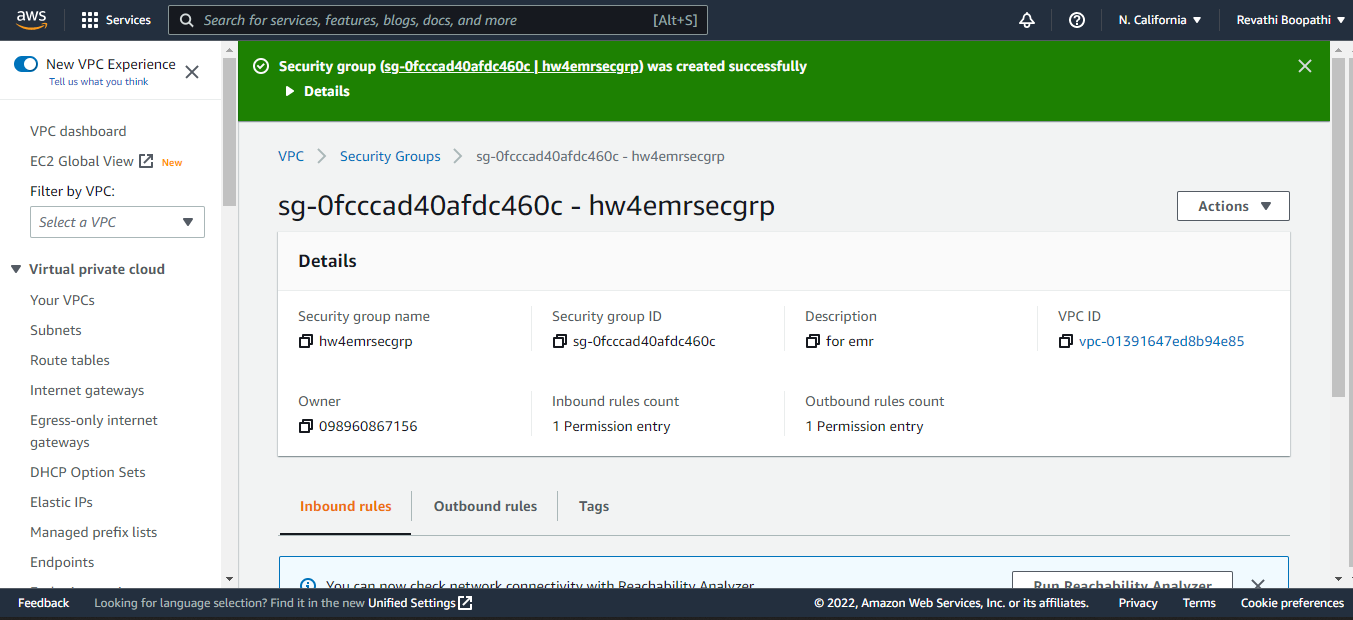






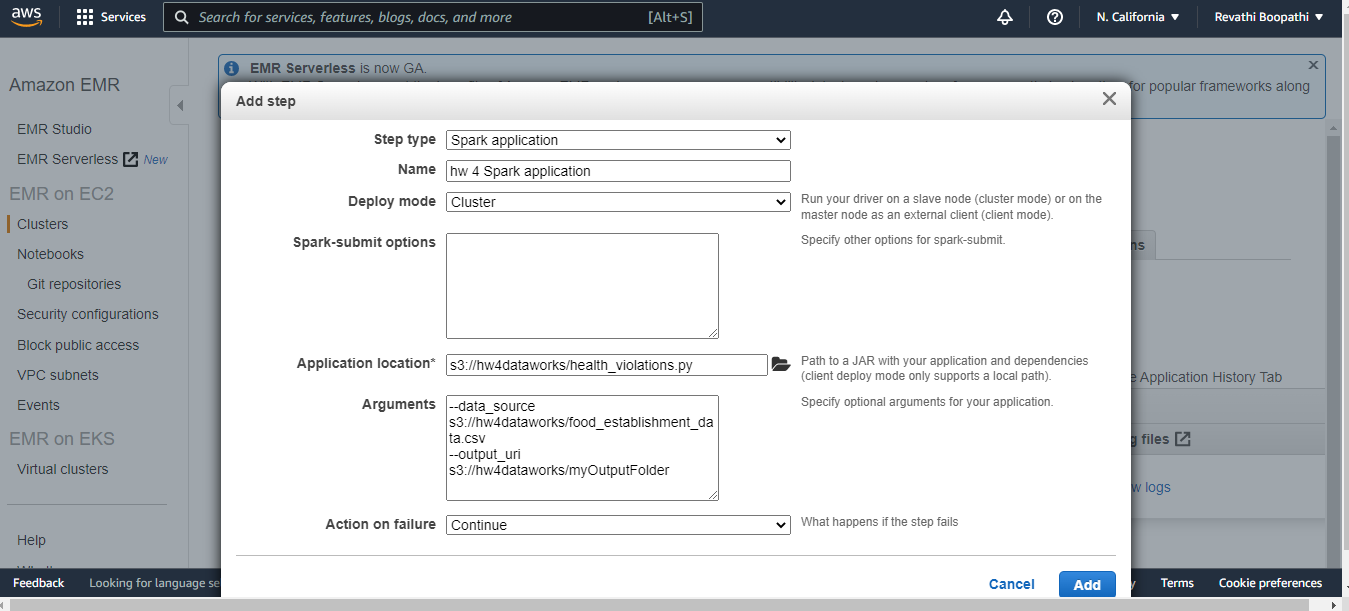
**Step 3: Allow SSH Access**

Setting inbound and outbound rules in order to allow access to the resources.

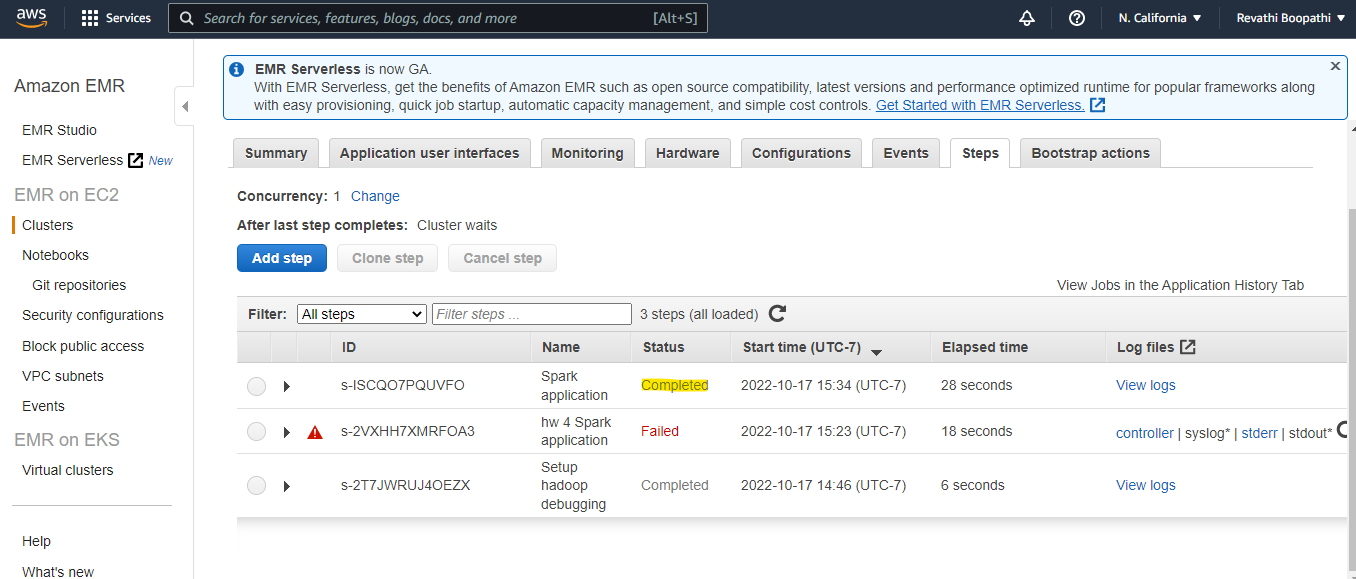


**Step 4 - Run a Hive Script to process data:**

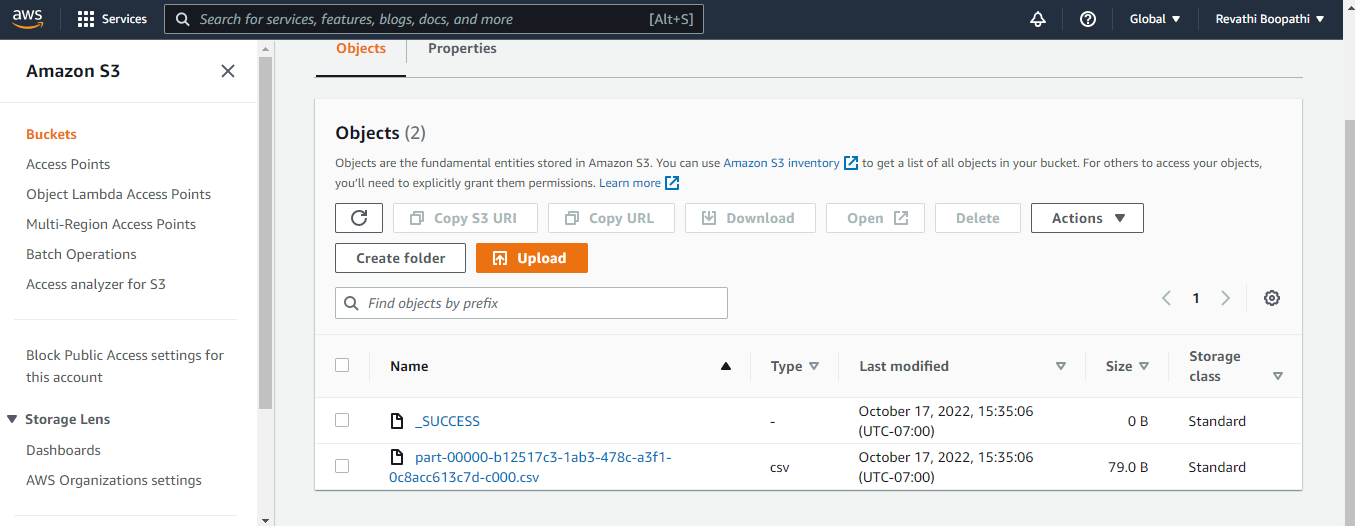
Submit work to cluster, which will process the data



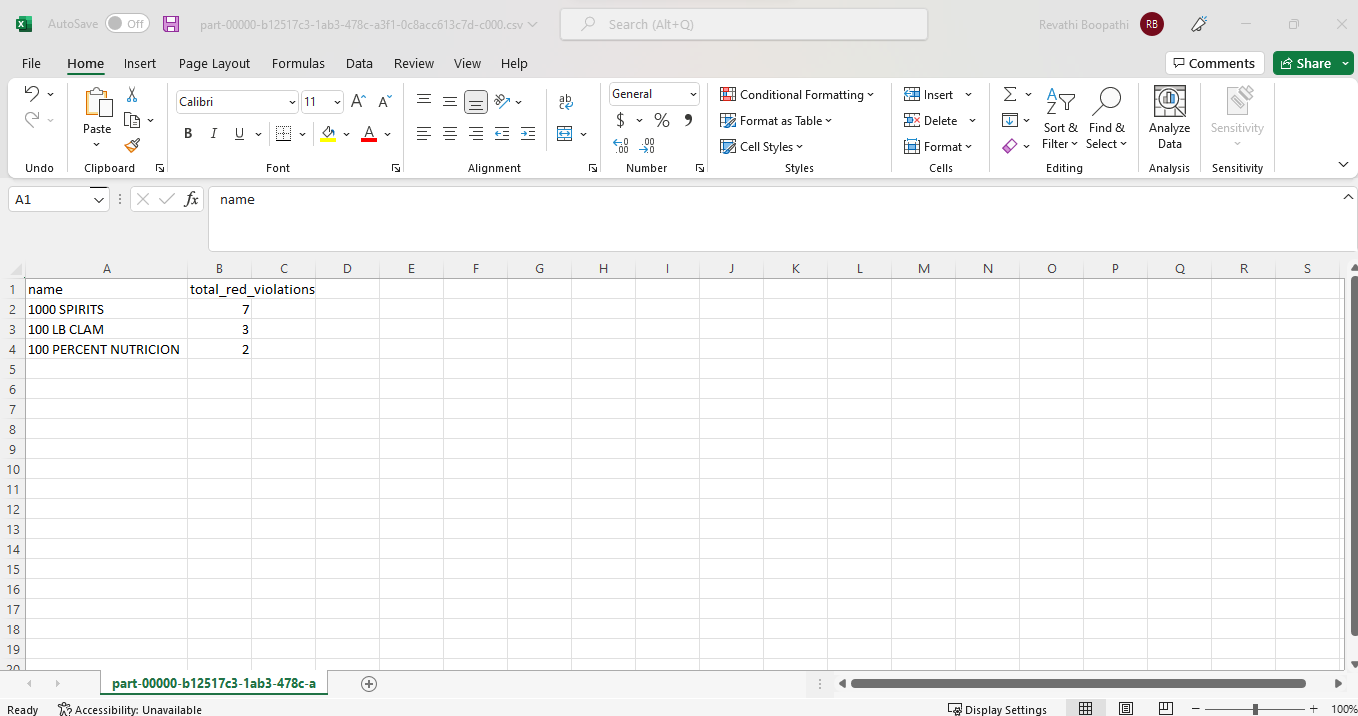
Check if the step completes successfully.



Ensure an output data is created in the folder called \_SUCCESS.



Validate the output that is created by downloading the file. (The below data shows the output from the sample of data in order to save space, so the number of records is less)



**Step 5: Clean up resources**

Terminate the cluster

