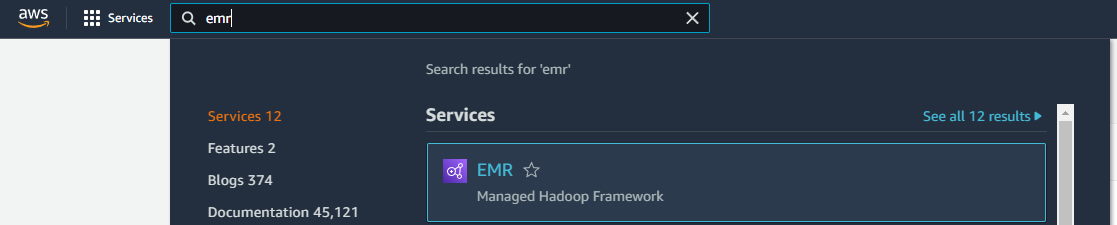
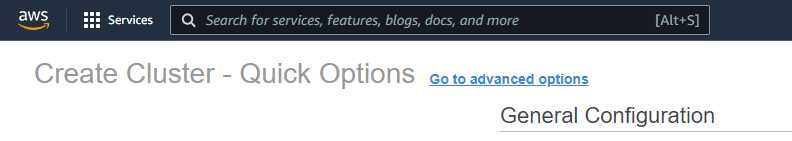
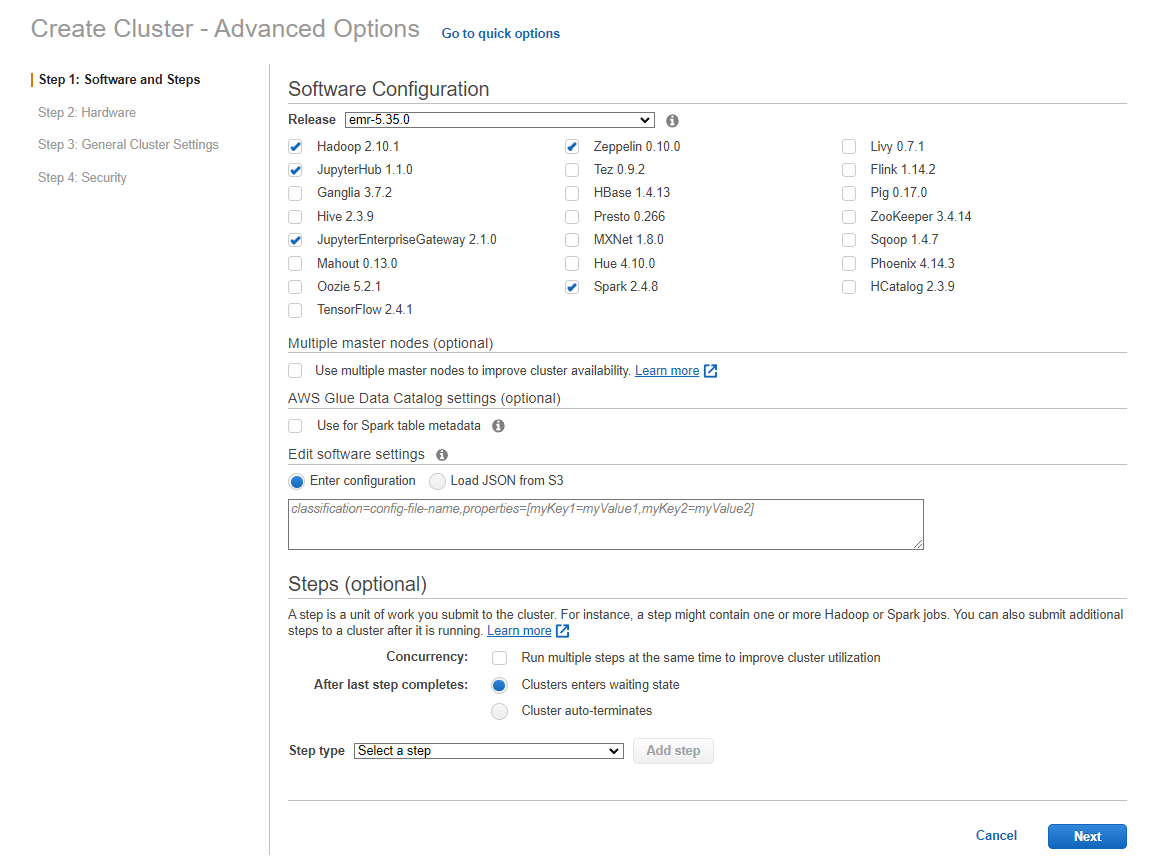
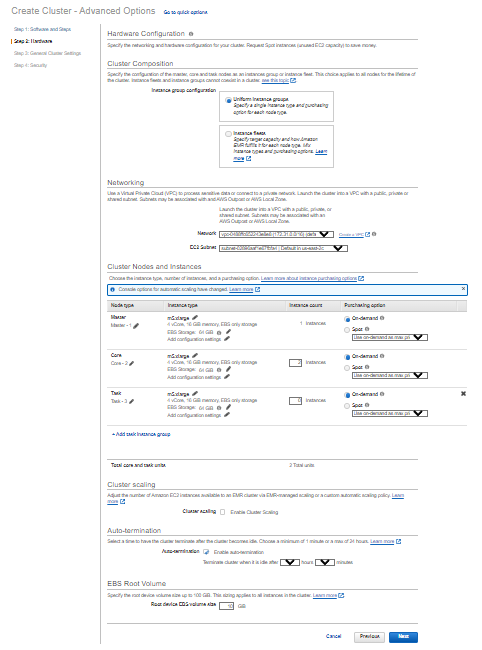
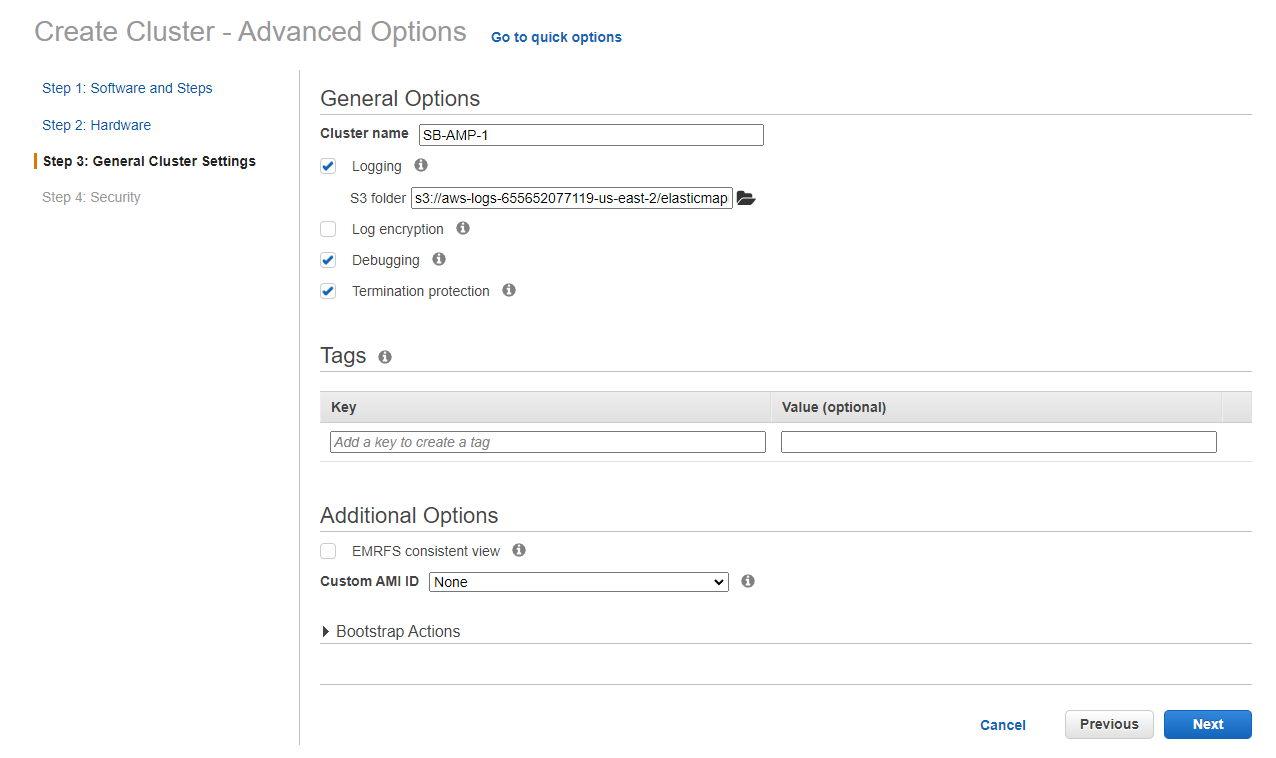
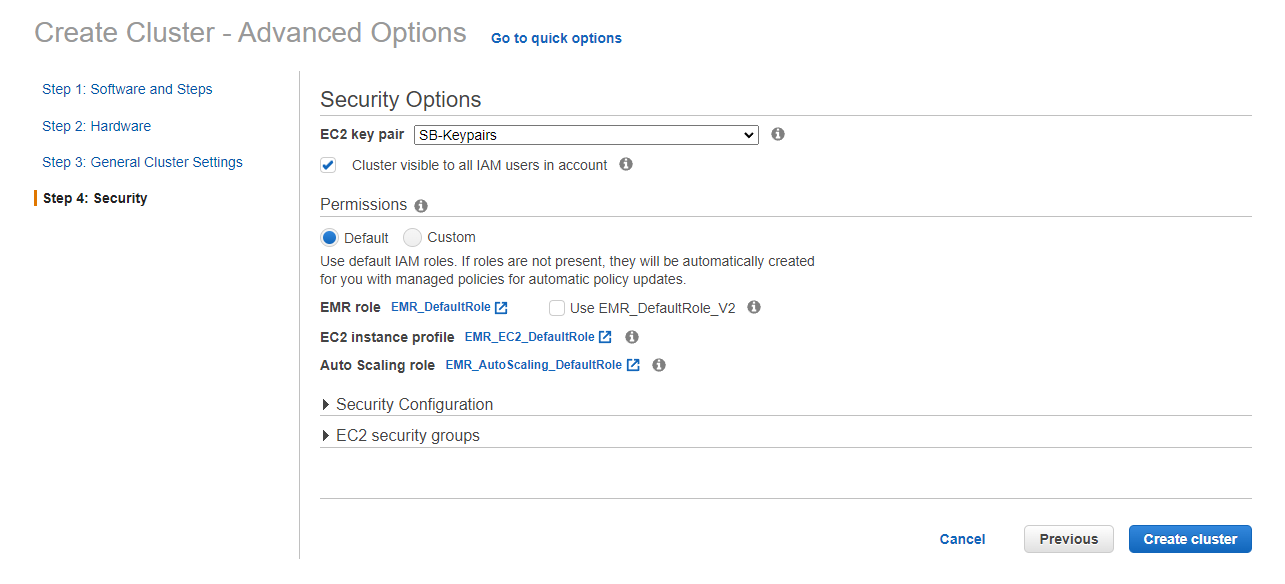
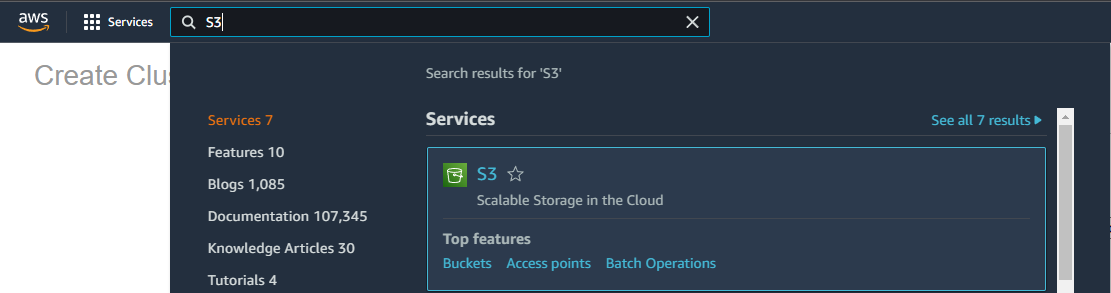
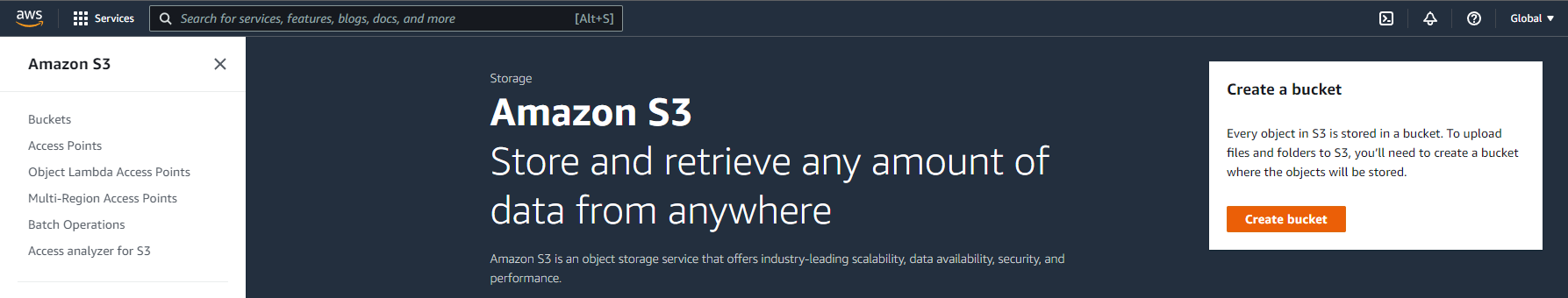
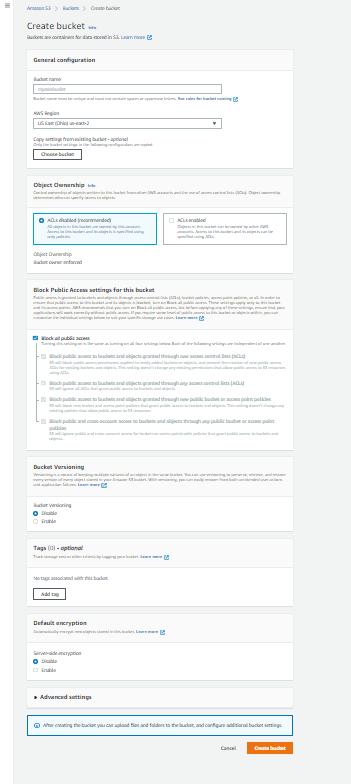
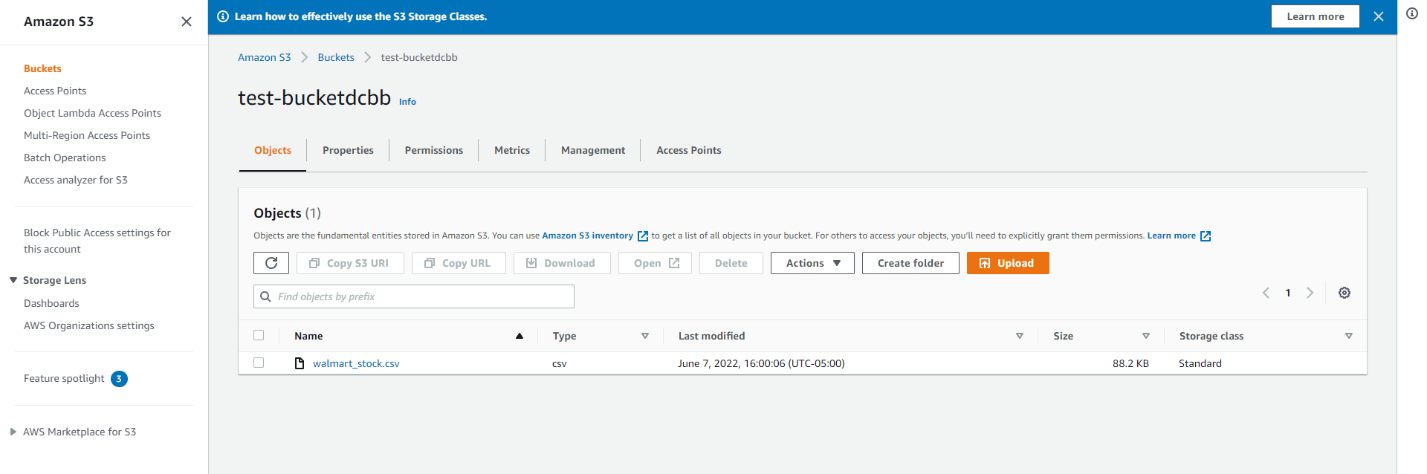
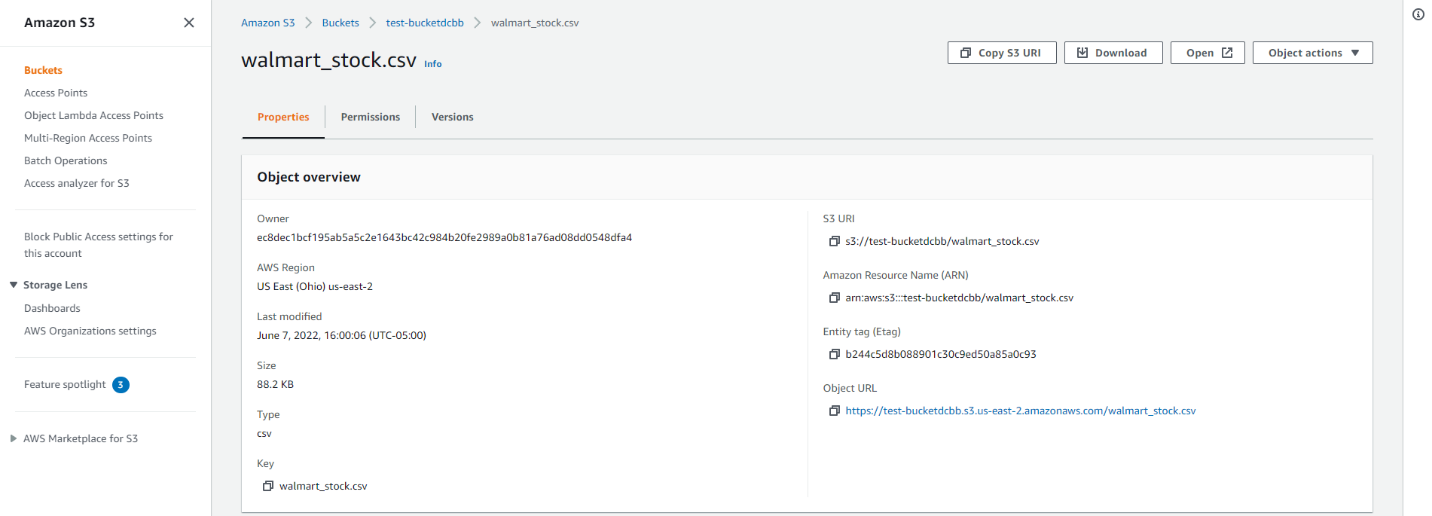
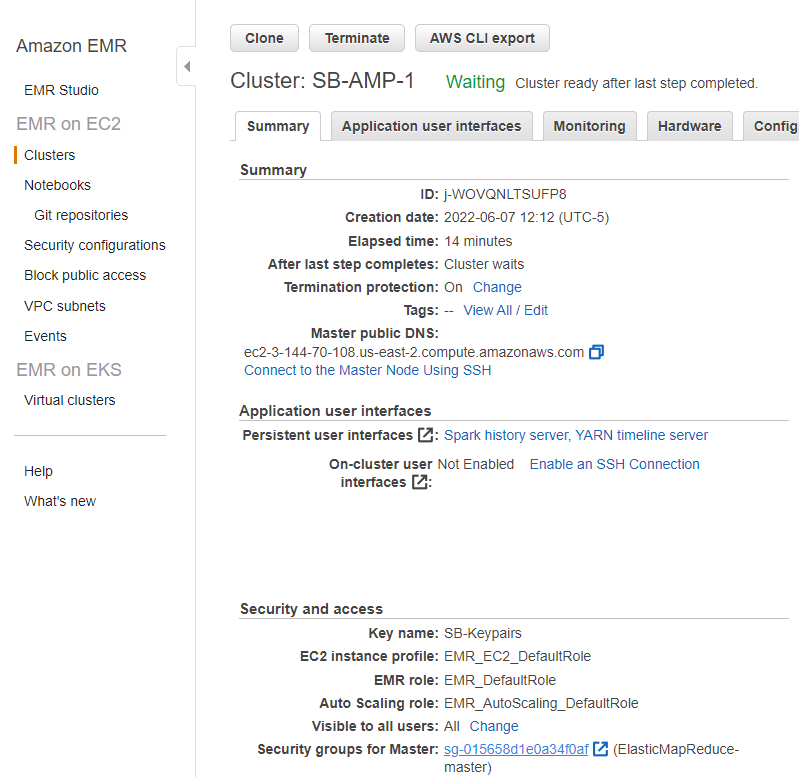
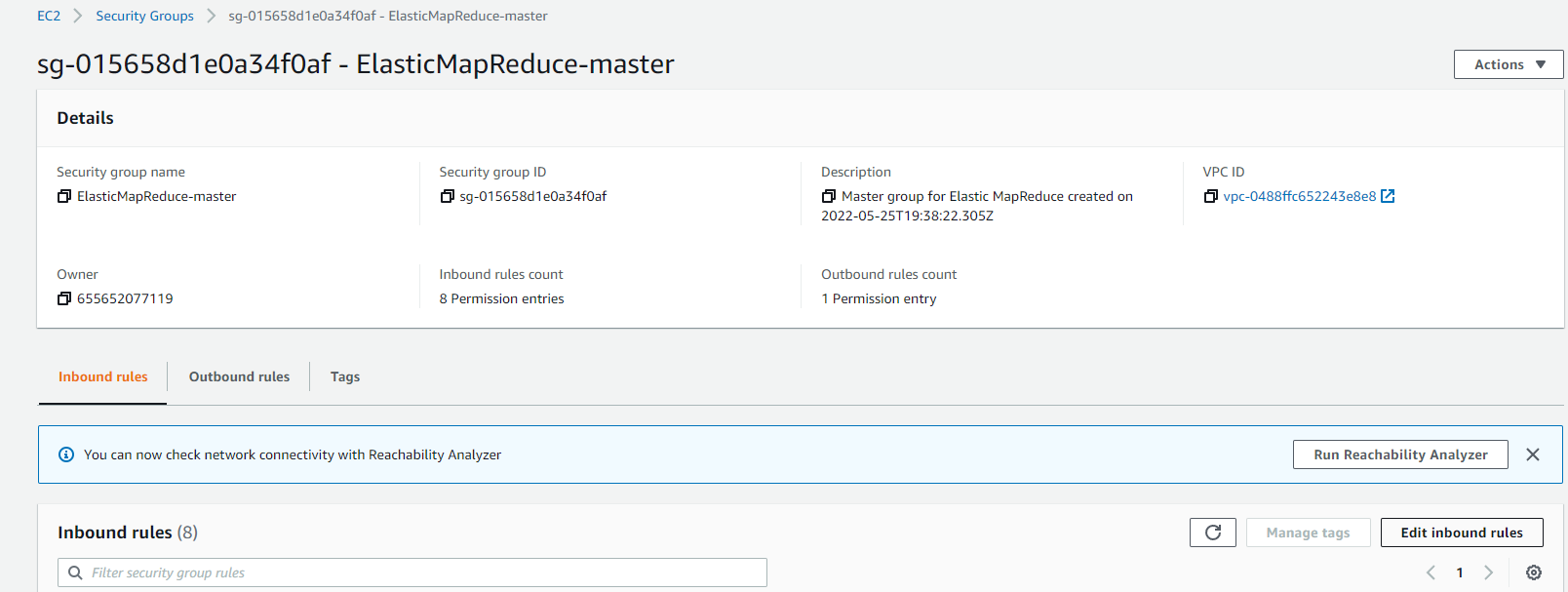
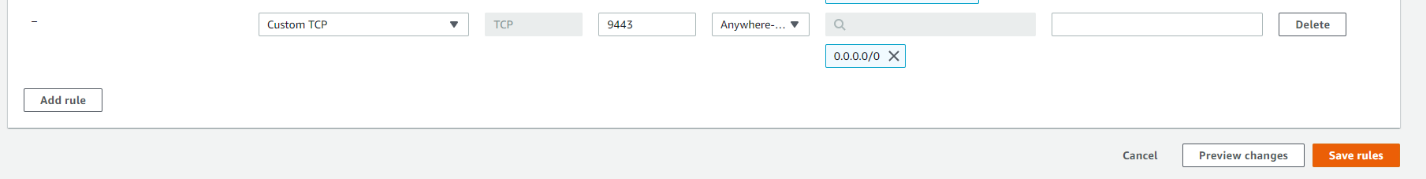
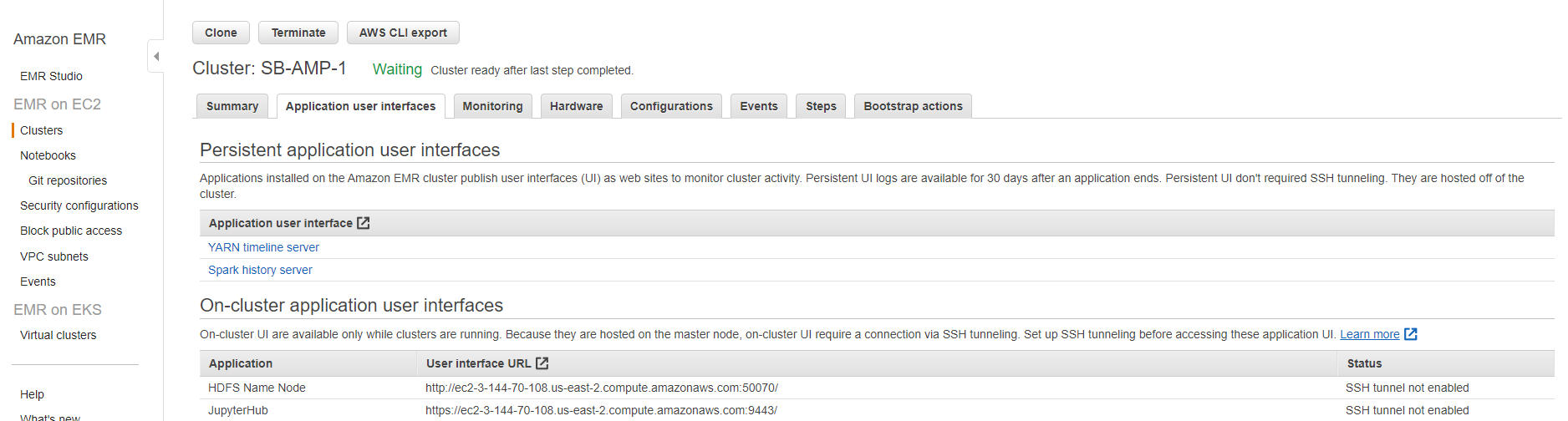
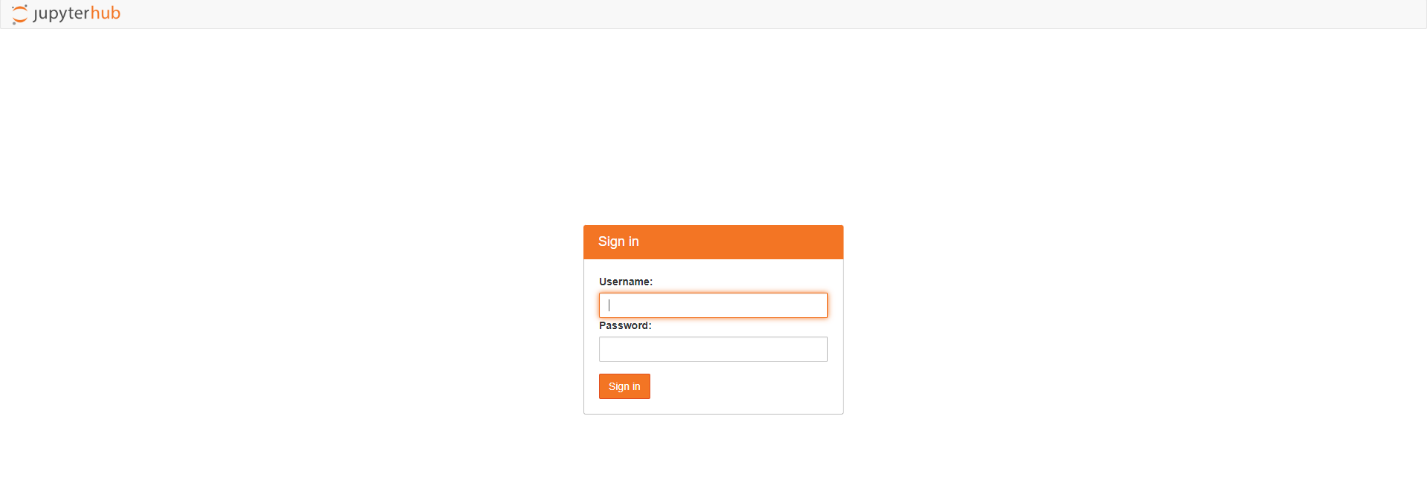
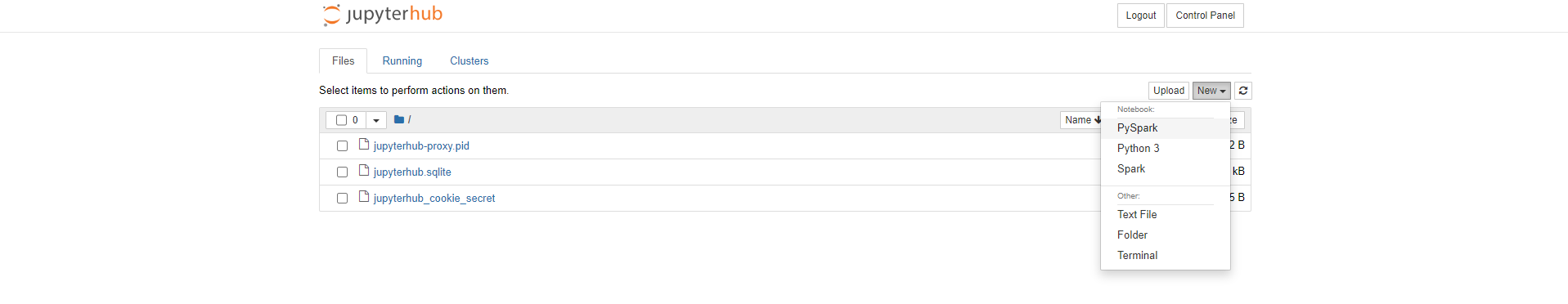
1. Once you’re signed into AWS search for EMR and click create a cluster. 
2. From the quick options page, click on go to advanced options at the top. 
3. From there, click on Hadoop, JupyterHub, JupyterEnterpriseGateway, Zeppelin, and Spark, then click next. 
4. Under hardware, leave everything as is and click next. 
5. Under General Cluster Settings, give your cluster a name and turn off termination protection. 
6. Under security, select an EC2 key pair, or make one by following the wizard, and leave everything else the same and click on the create cluster button.
7. While the cluster is starting up, search for S3 to create a new bucket. 
8. Click on the create a bucket button, or the buckets link in the left navigation.
9. Create a new bucket with a unique name and turn the default encryption to enable.
10. Once in the bucket, add the csv file you will be using in spark. 
11. Click the file name and then copy the S3 URI. This will be needed to load the file into the jupyter notebook with Spark.
12. Go back to the cluster once it is started and click the security group for master link appears then click on it. 
13. Once here, click on the edit inbound rules button. 
14. Click add rule at the bottom of the page. In the port box type in 9443 and add 0.0.0.0/0 after clicking on the box with the search icon, then click the save rules button. 
15. Back in your cluster, copy and paste the jupyterhub link into a new tab. 
16. Once here, use the default login of jovyan and password jupyter. 
17. Once logged in, you can create or upload a jupyter notebook. Using the S3 URI link from earlier, you can connect to your data using spark. 
18. The important commands that I used to start the session and import the data were:

from pyspark.sql import SparkSession

from pyspark.sql.functions import col,round,max,min,avg,count,year,month

df = spark.read.options(delimiter=',').csv('URI GOES HERE',inferSchema=True,header=True)