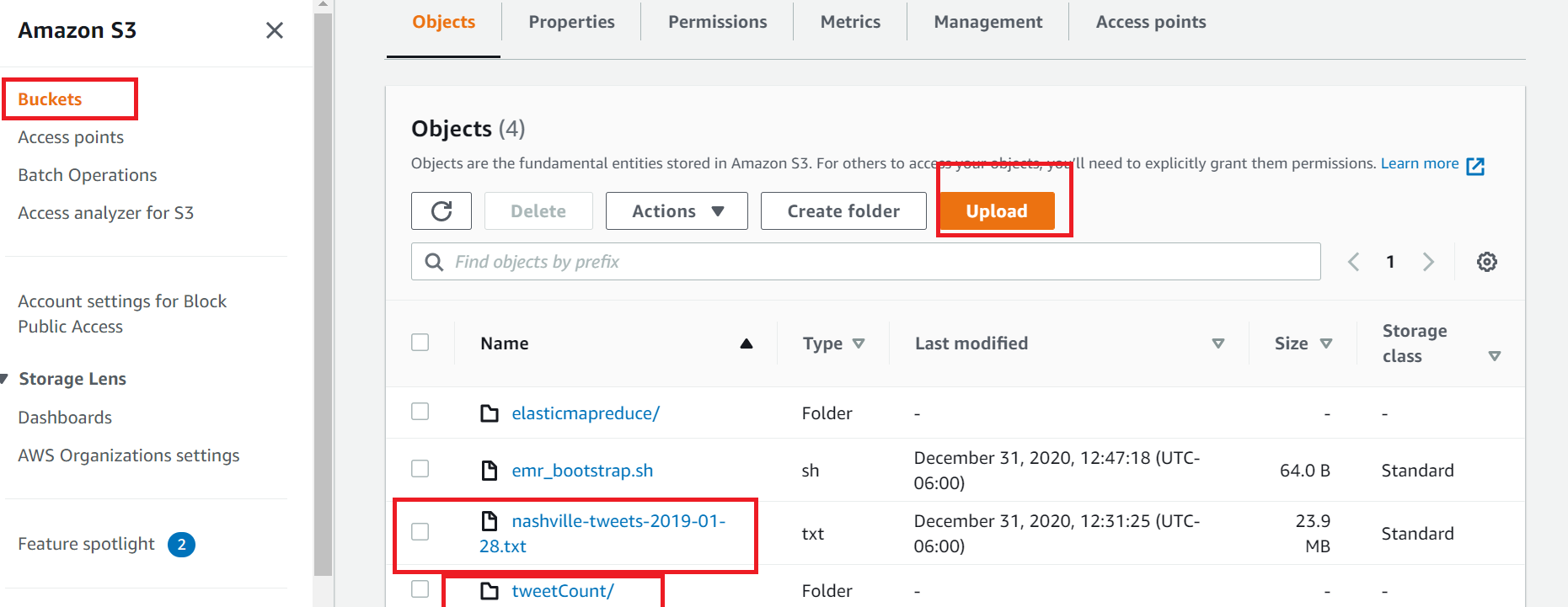
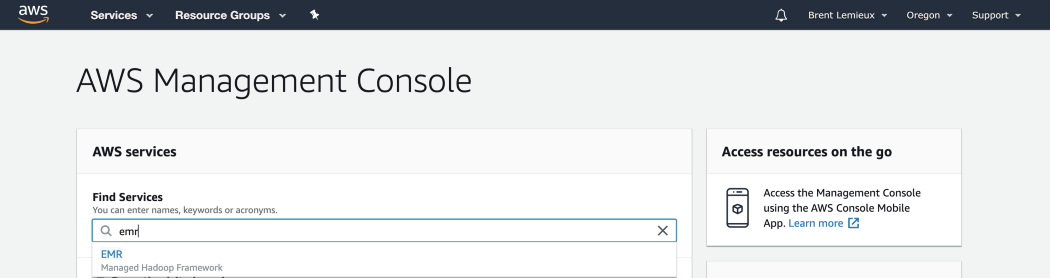
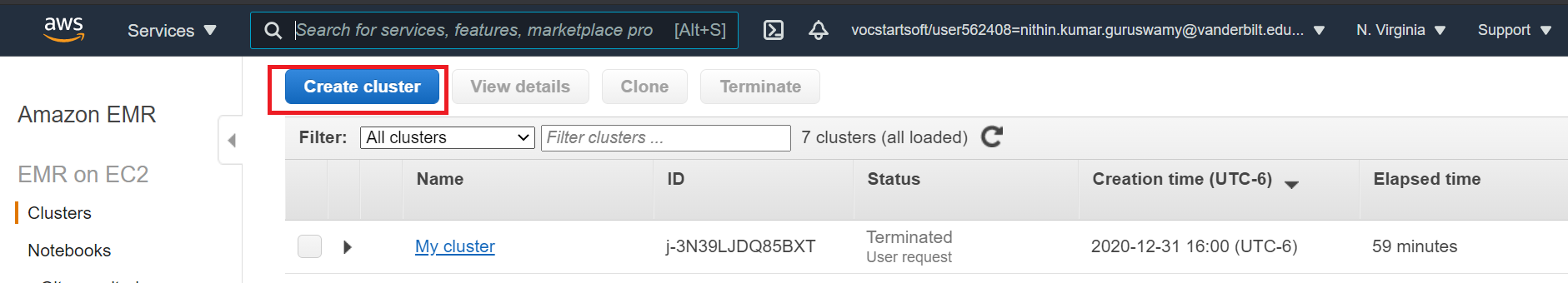
PySpark on AWS EMR

* Create a new S3 bucket and upload the bigdata nashville tweet file into the newly created S3 bucket.

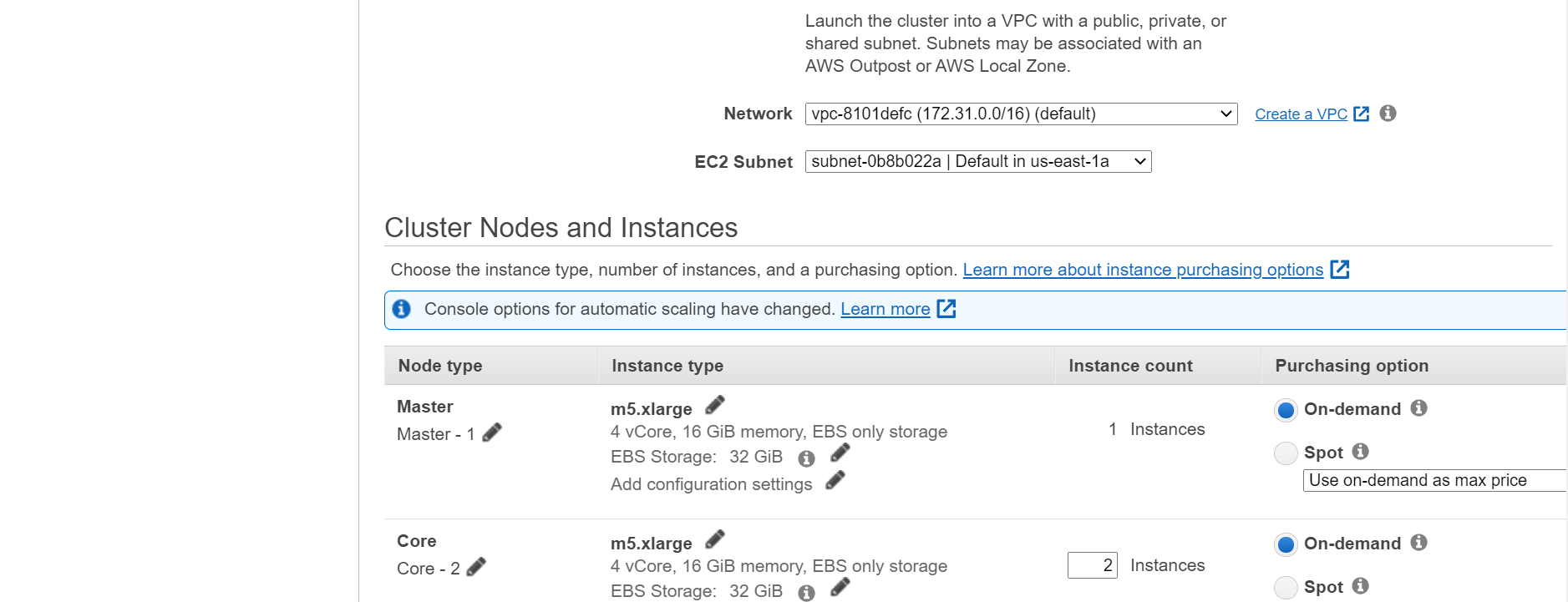


* Create a new AWS EMR cluster to run the pyspark example

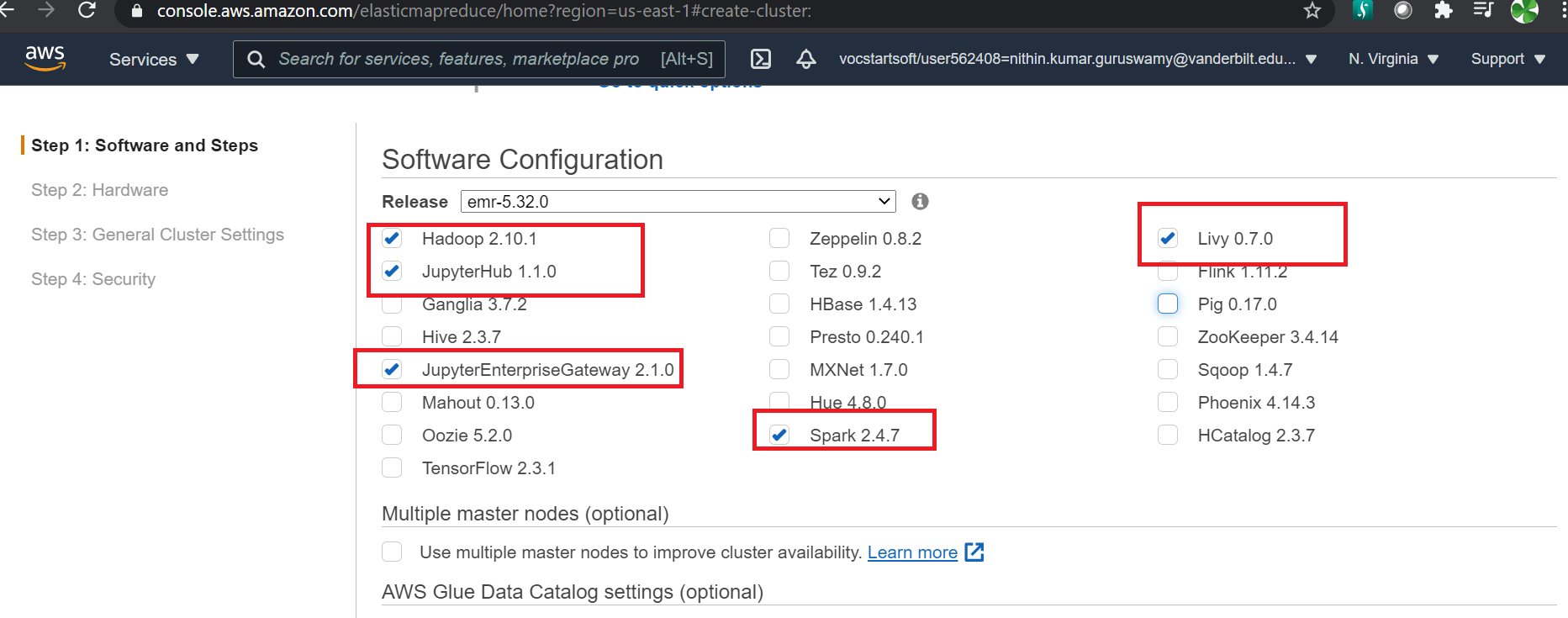




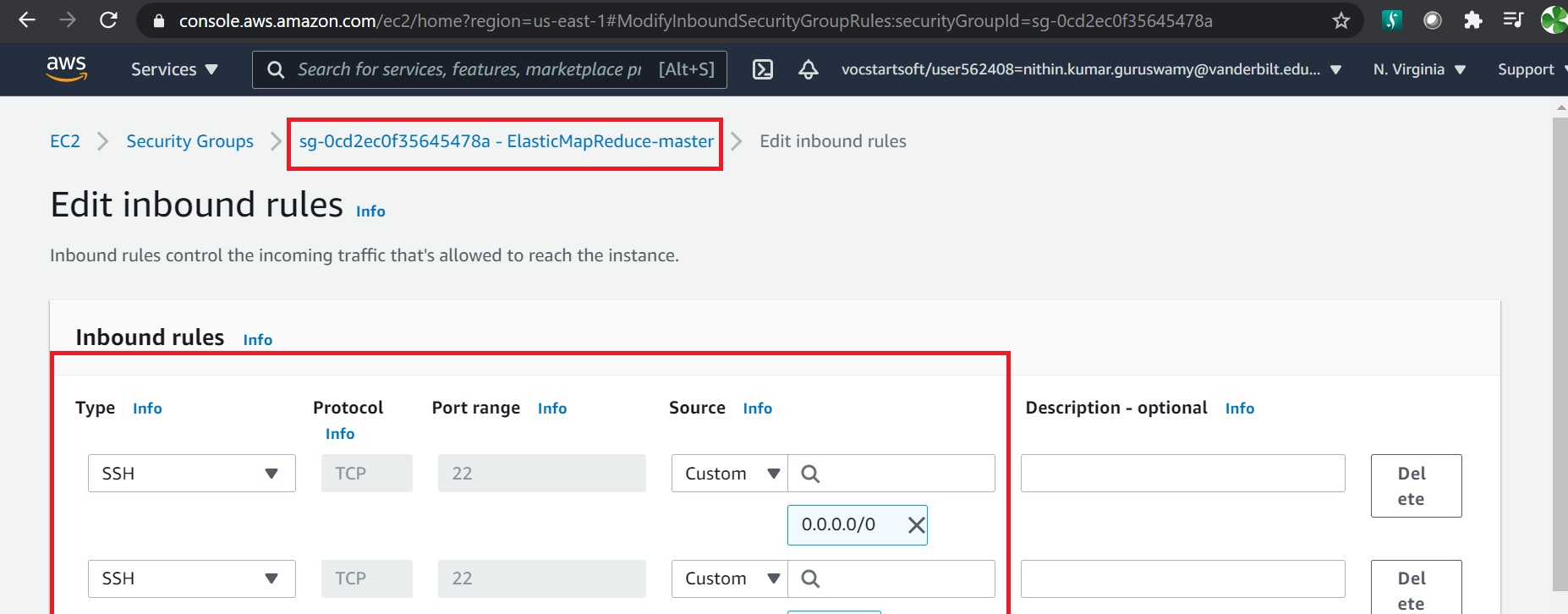
* Leave some of the cluster size to default values



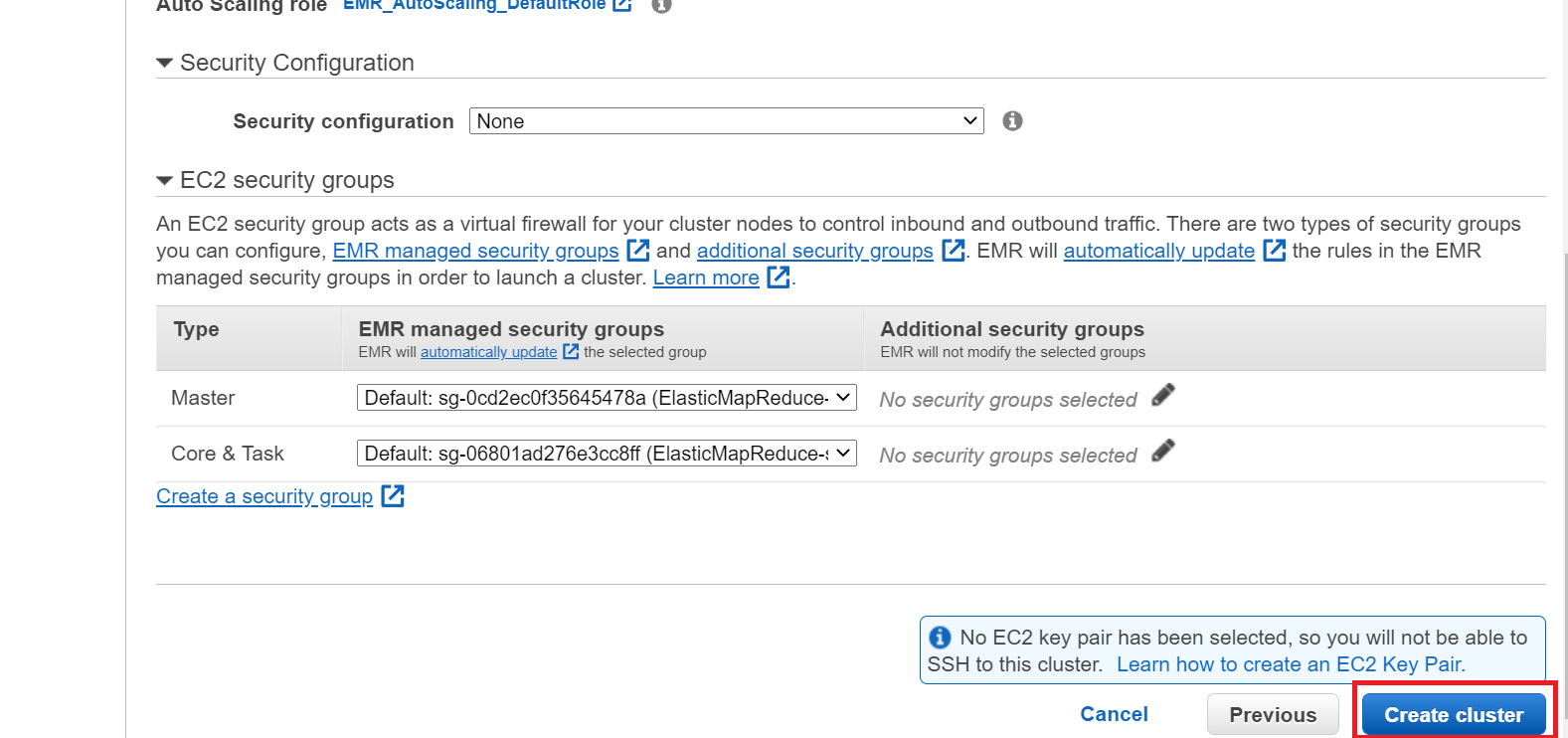
* Select the following software packages to be installed on the EMR clusters



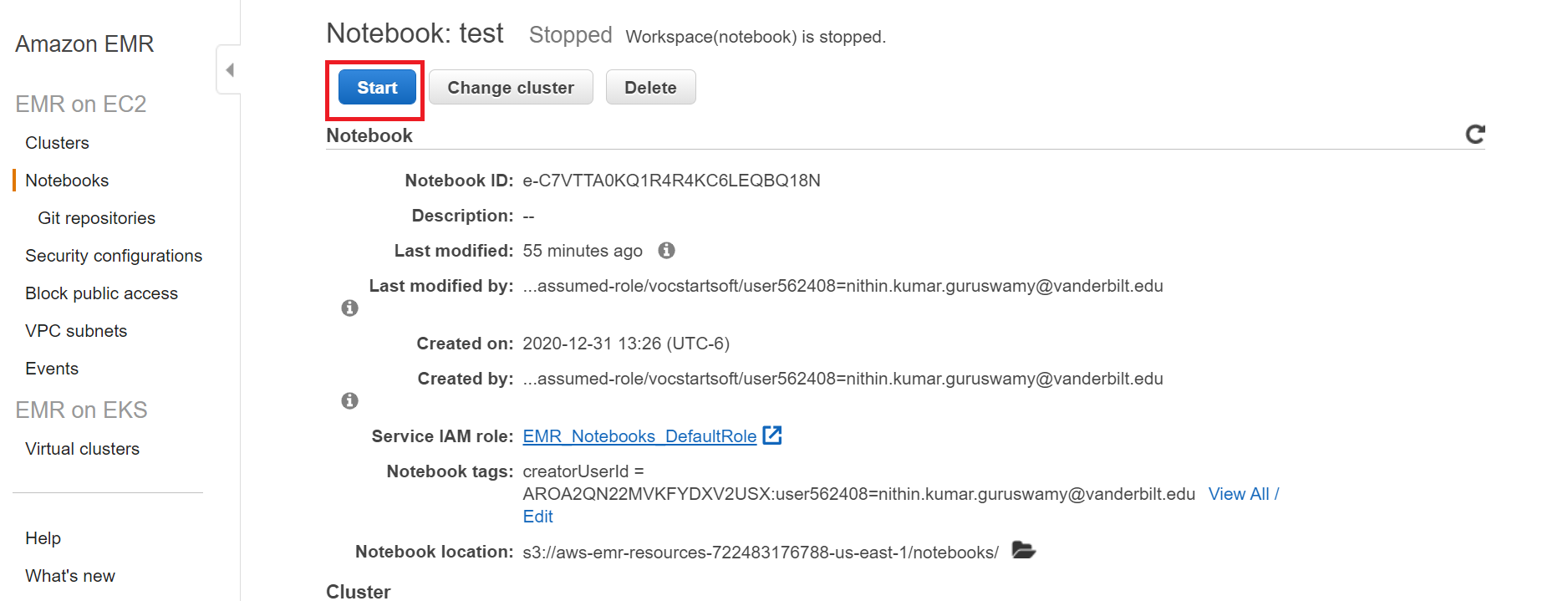
Before creating the clusters make sure the security groups of slave and master node have only port 22 open



* Create the cluster

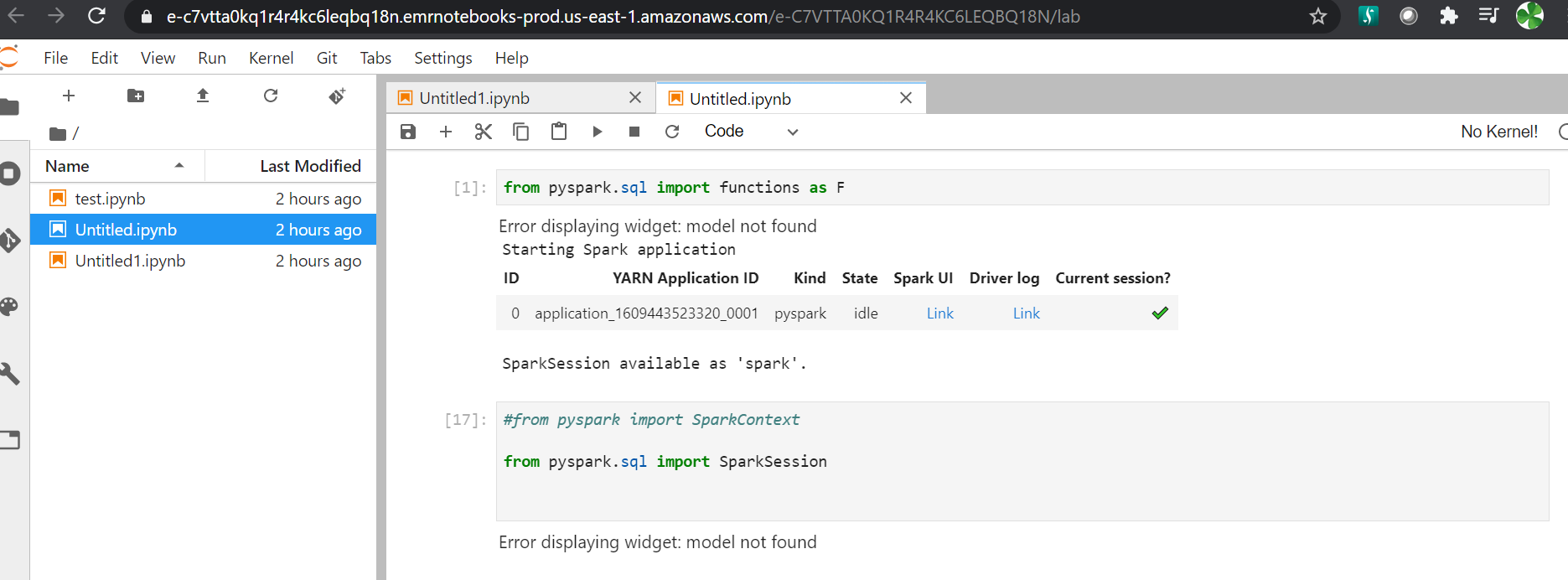


* Create the notebook to run on the cluster

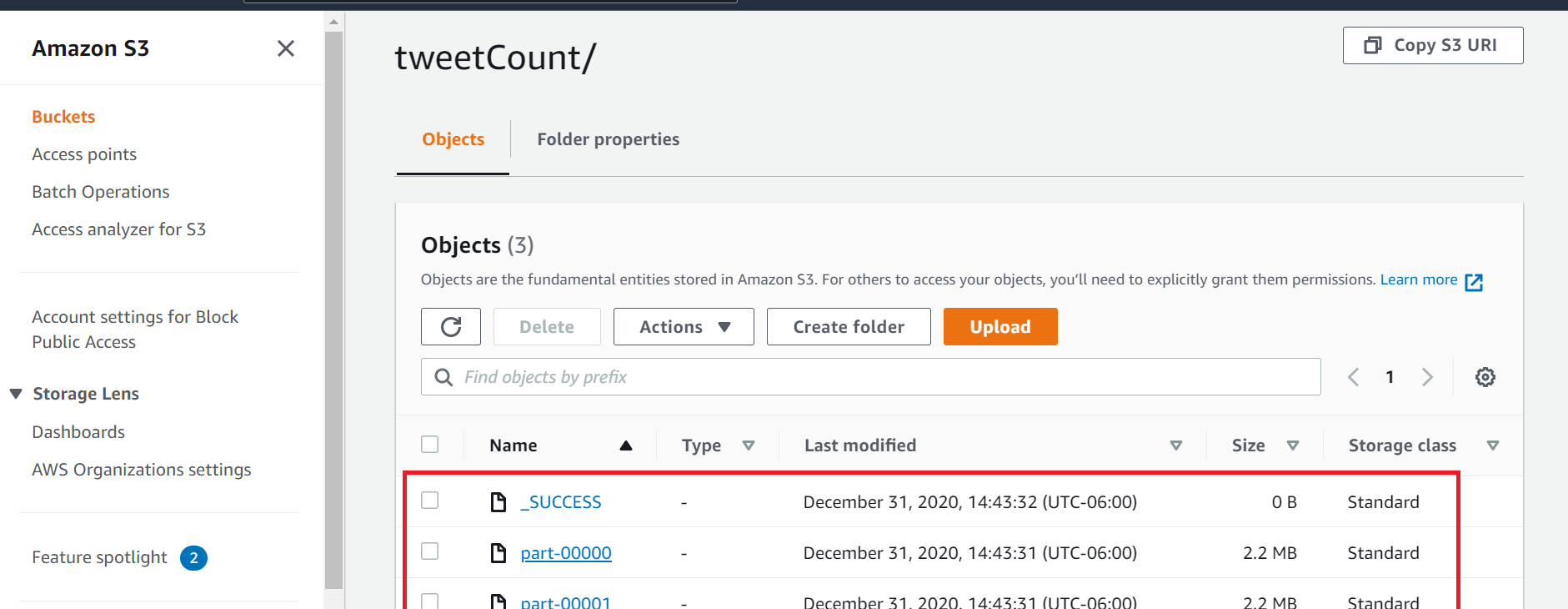


* Start the jupyter notebook and run the tweet word count example

<https://colab.research.google.com/drive/1i11lBzJI_MDkLdKcQrHOGtHvtxTpENnp#scrollTo=lU-0ARzkvgDg>



* The solution will be in the S3 bucket



* Terminate the EMR cluster , Notebook and delete all S3 contents to prevent from being charged.