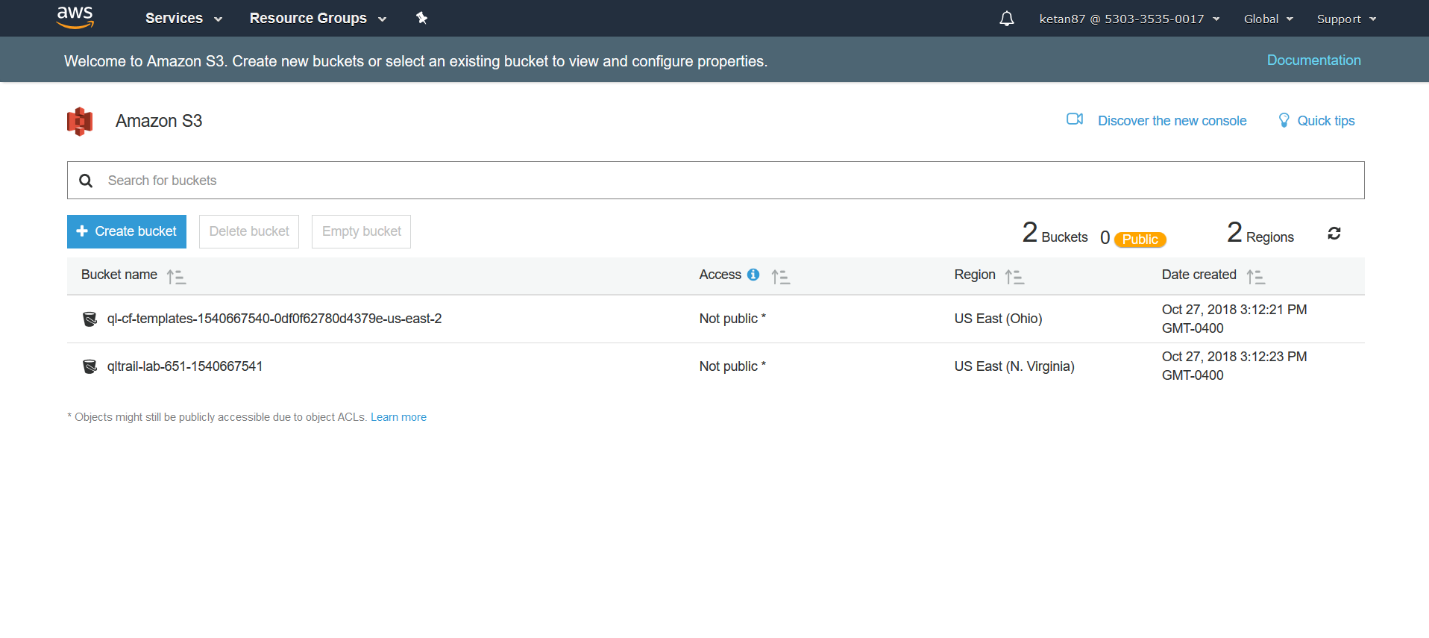
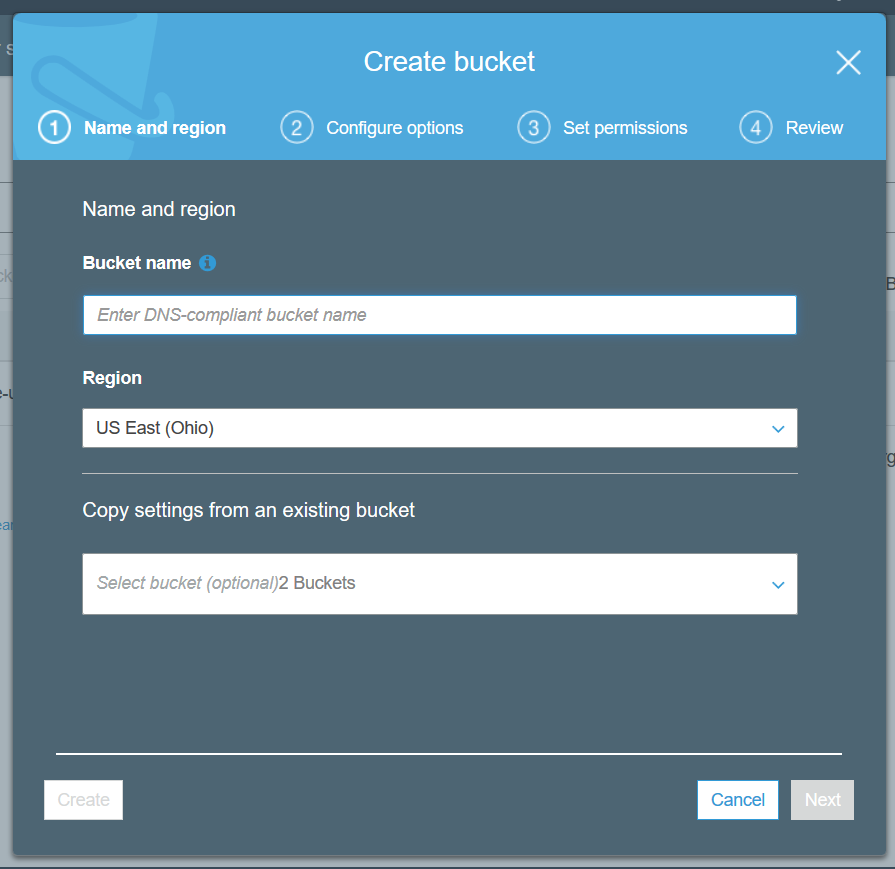
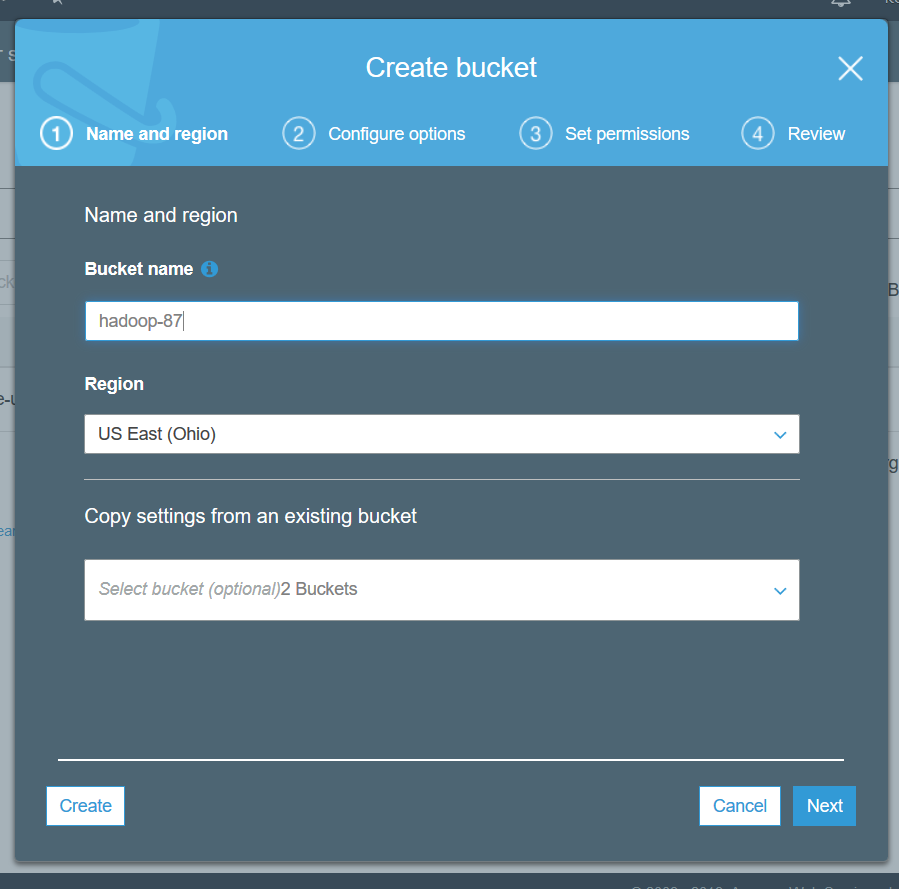
Analyze Big Data with Hadoop

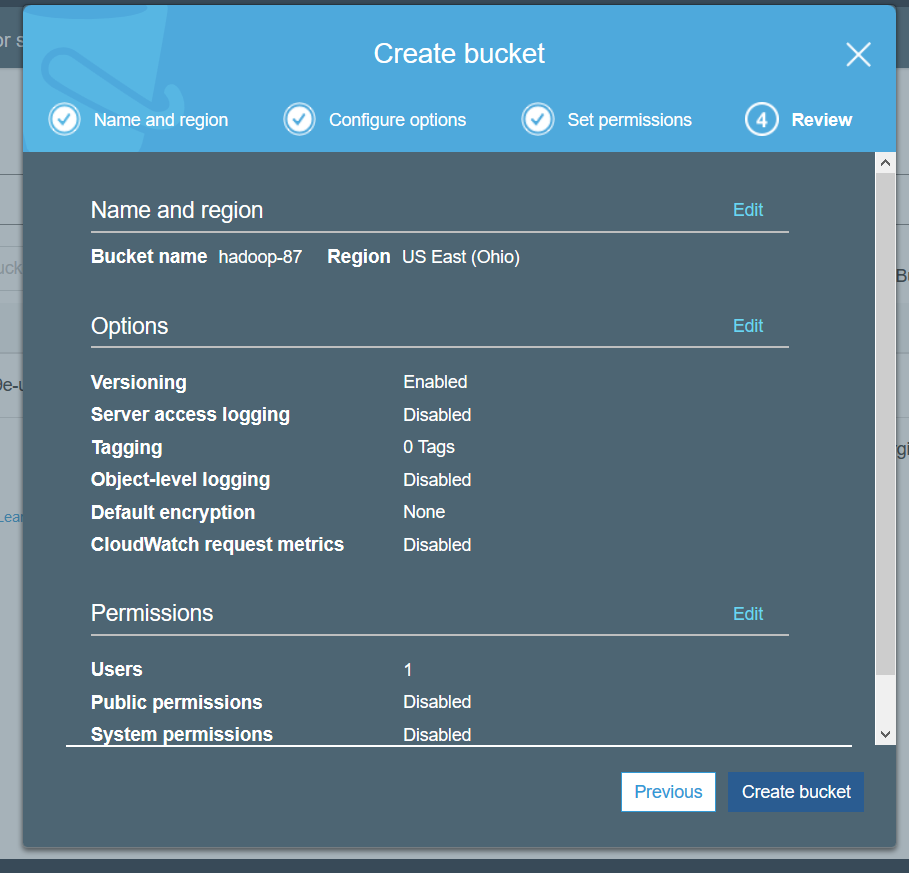
1. Visit <https://awseducate.qwiklabs.com/focuses/19?parent=catalog> and create a few account and start the lab
2. Click Open console which opens the AWS Console
3. Create an S3 Bucket
   1. Open S3 service console

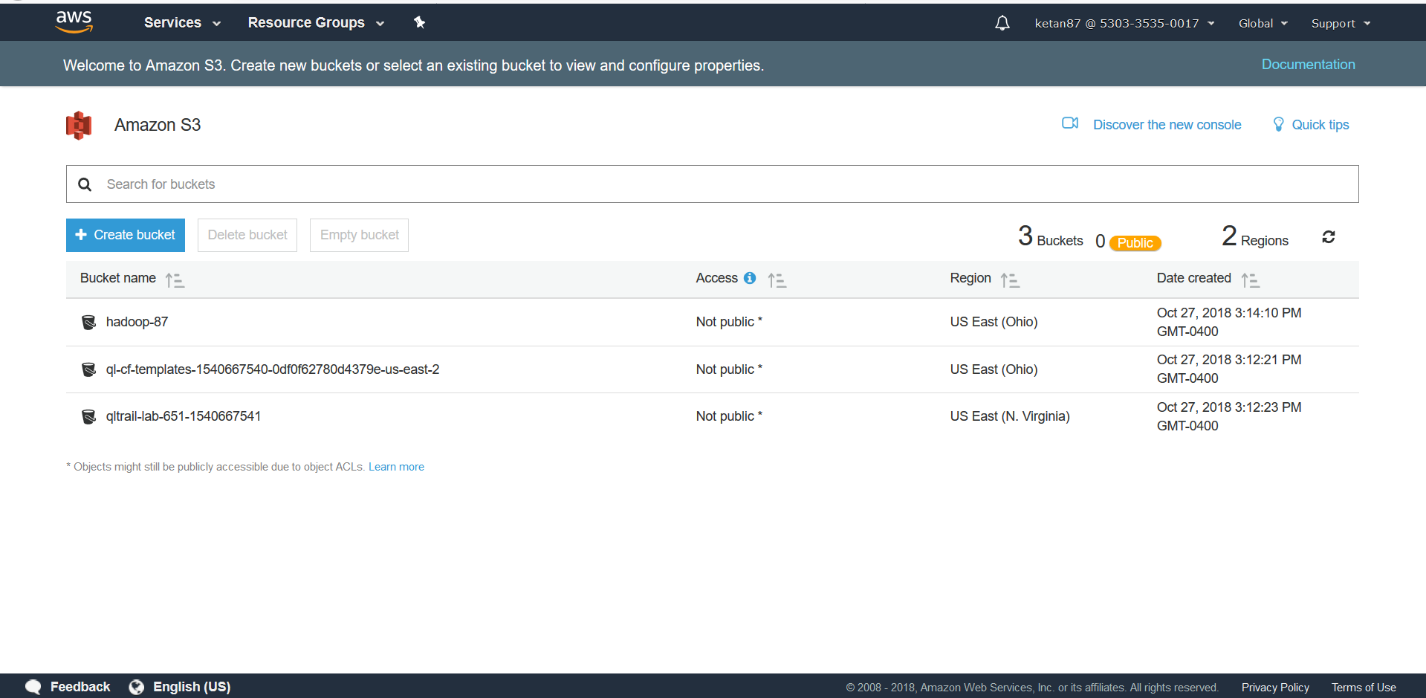


* 1. Click on create bucket and enter bucket name as Hadoop-<number>

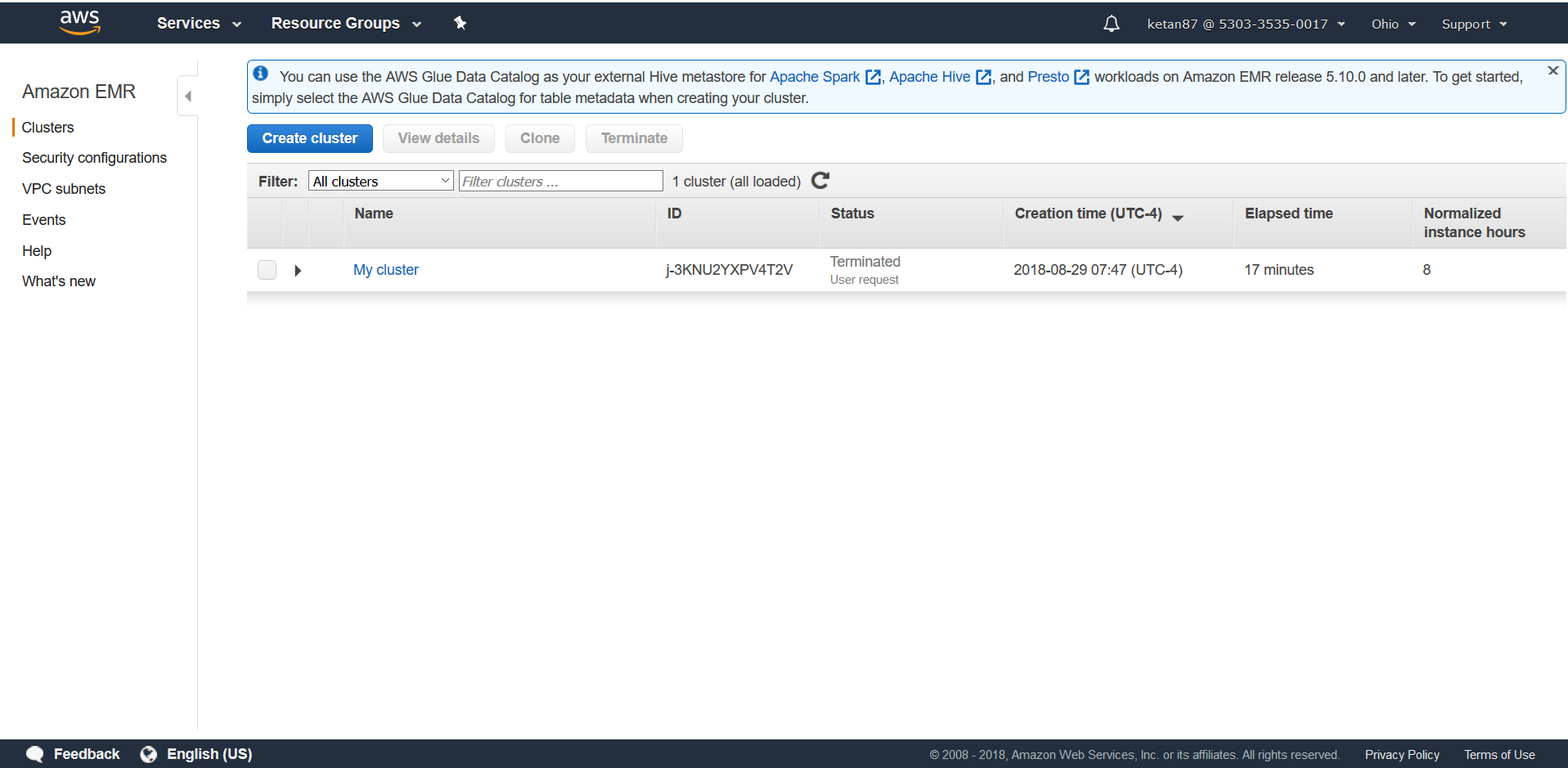


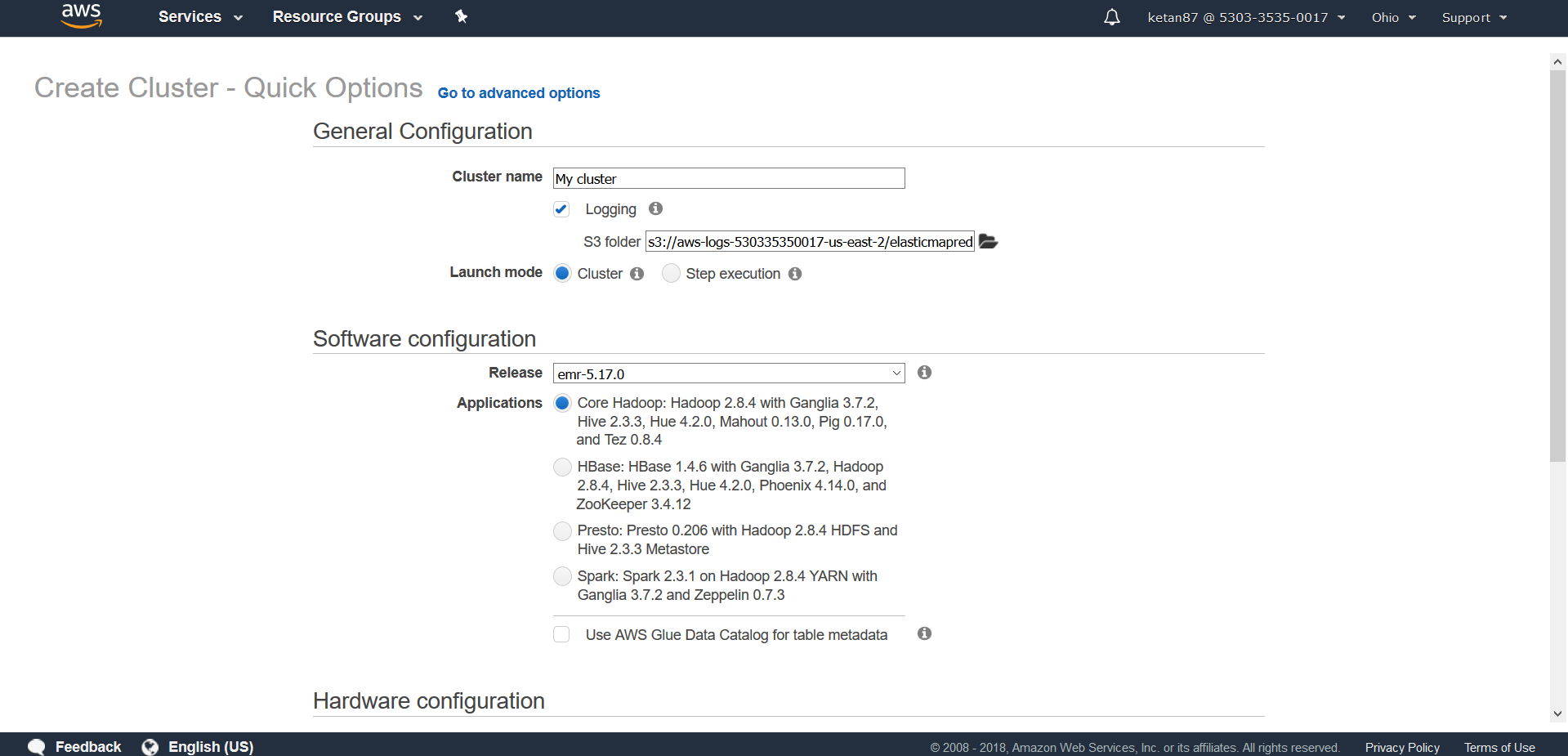


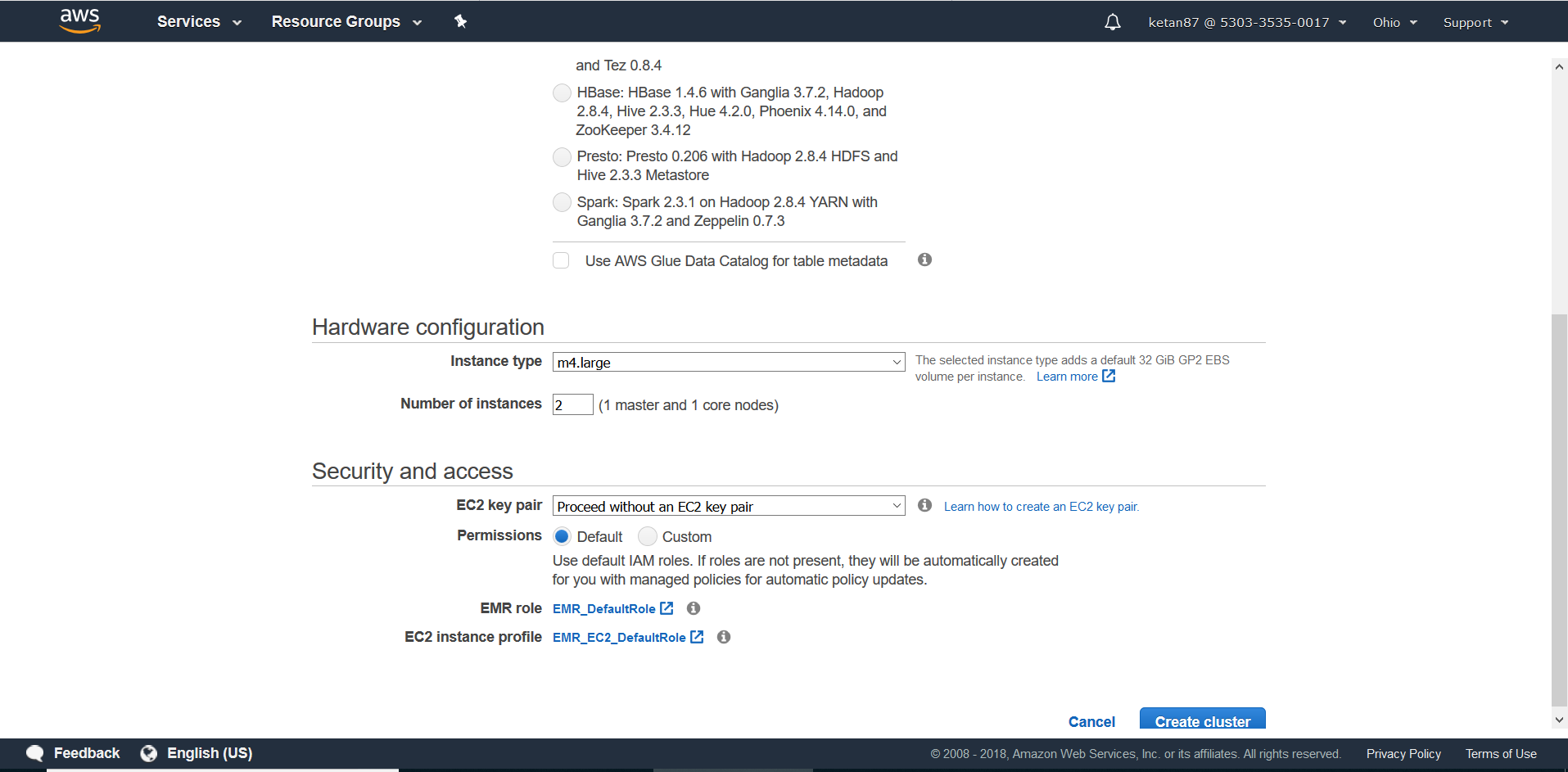


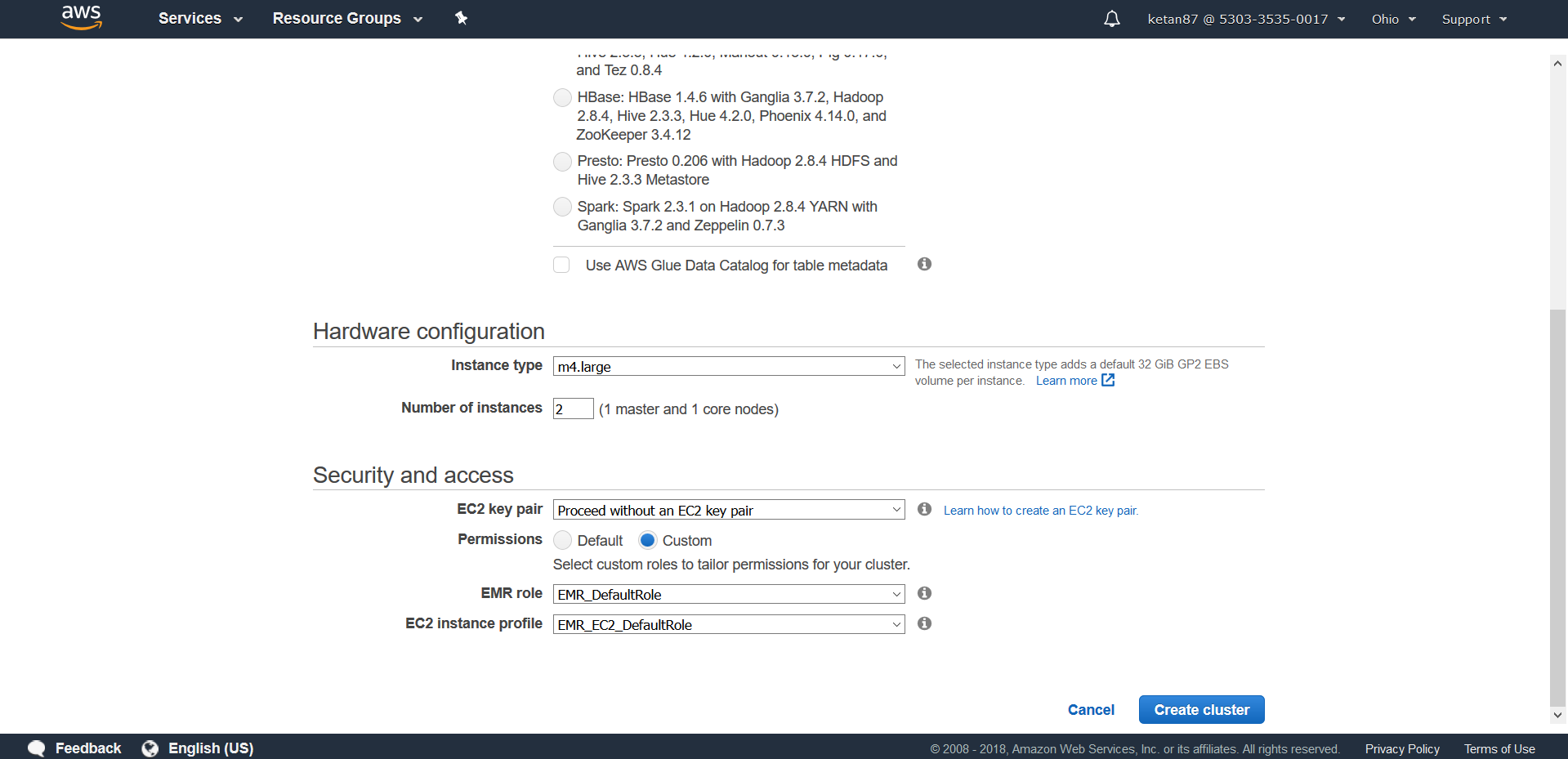


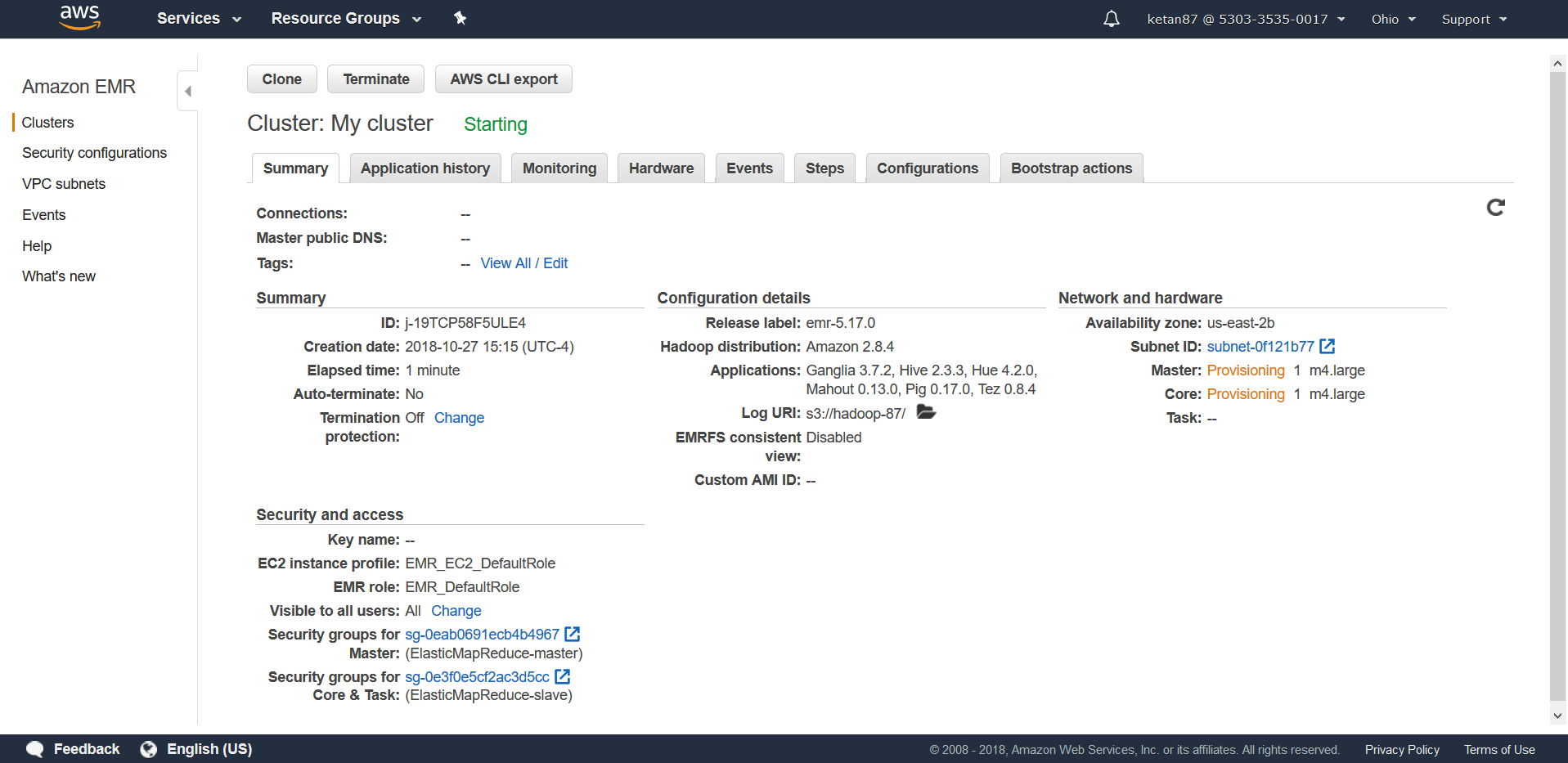
1. Launch an Amazon EMR Cluster
   1. Launch an Amazon EMR cluster with Cluster Name – “My cluster” and S3 folder as the S3 bucket created above
   2. Select m4.large as instance type and number of instances as 2 with one as master and another as slave
   3. In security, configure EC2 without key pair. Permissions: Custom; EMR Role: EMR\_DefaultRole; EC2 instance profile: EMR\_EC2\_Default\_Role
   4. Launch Cluster







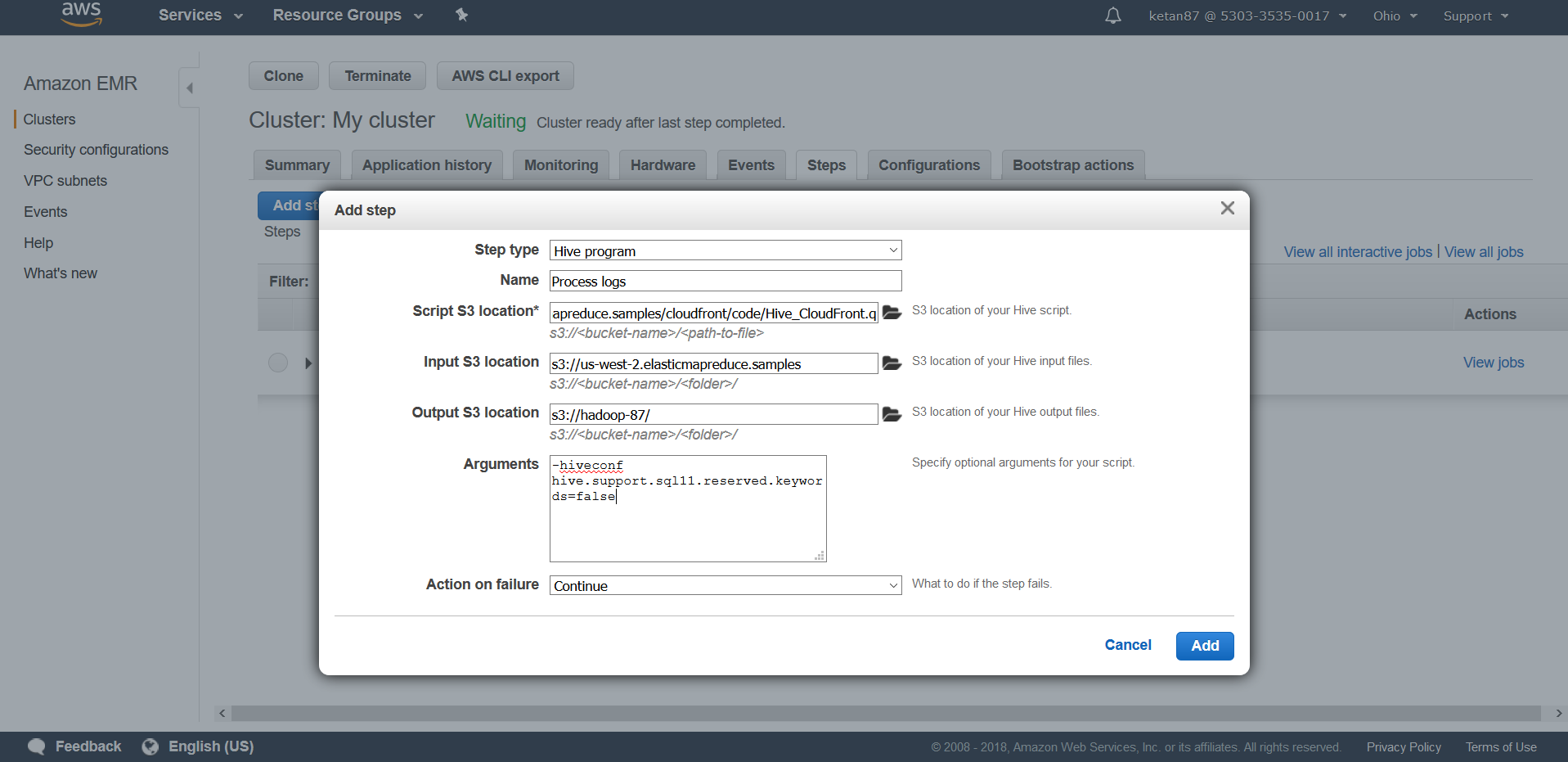


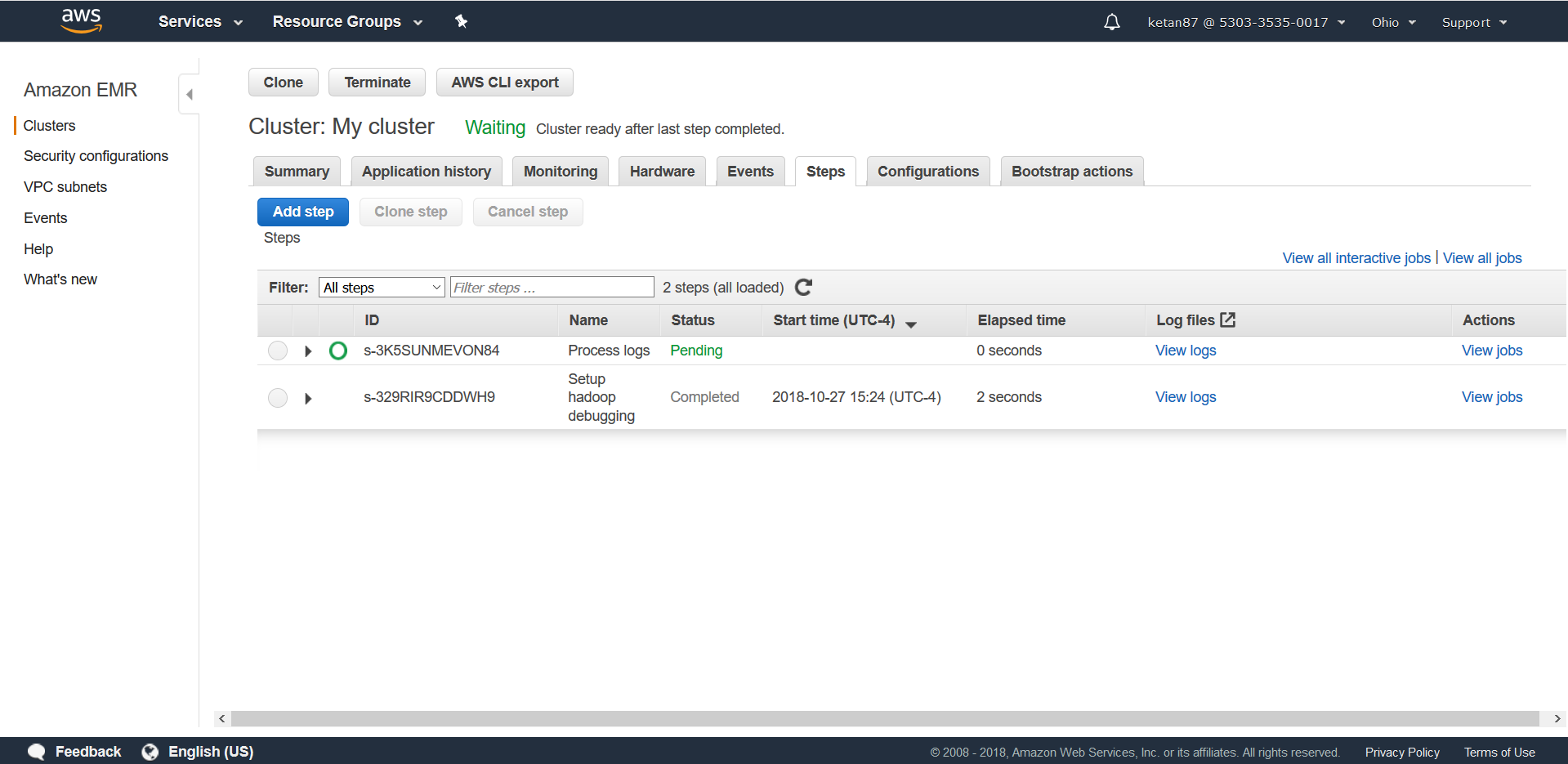


1. Data to be processed

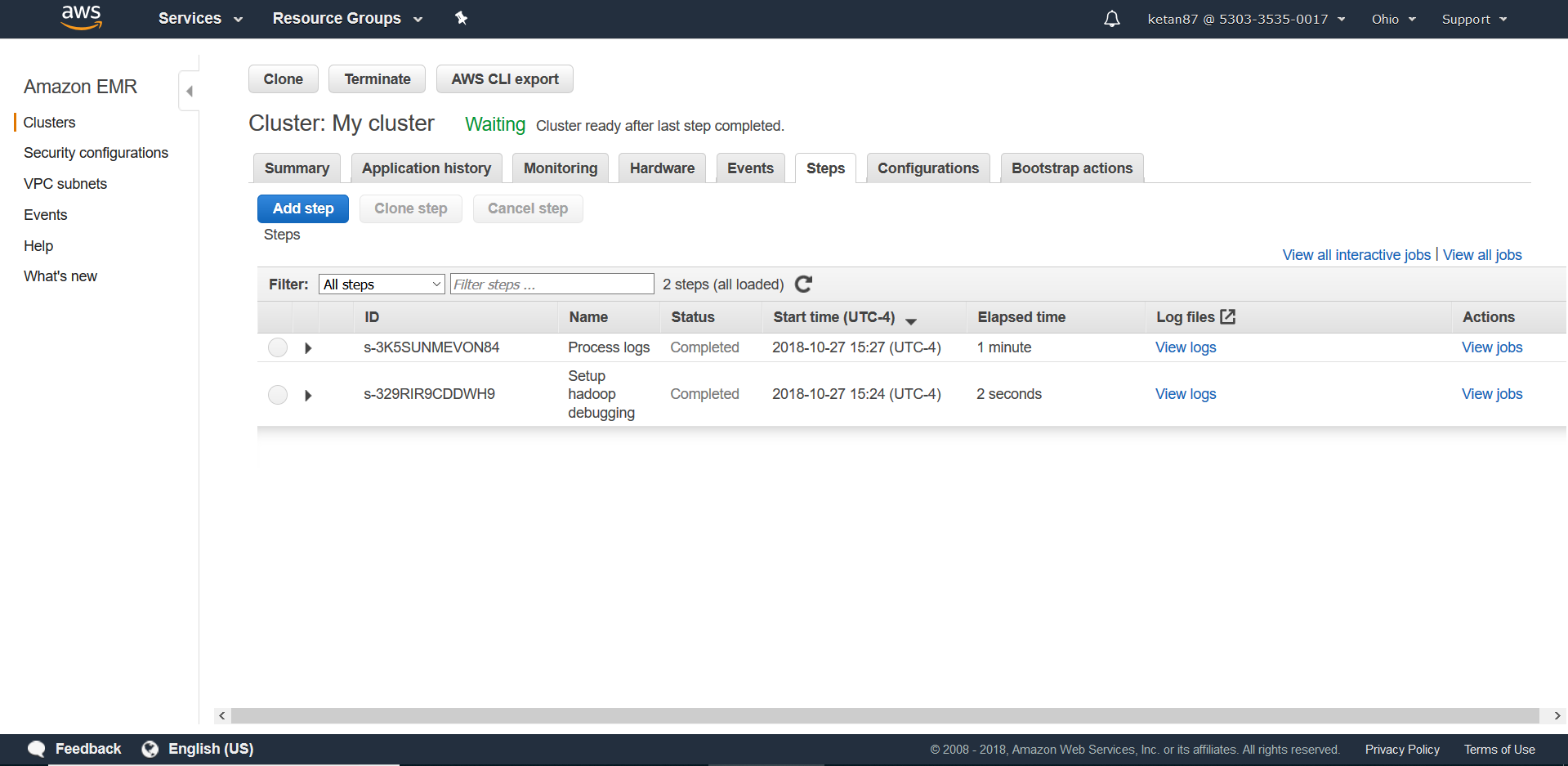
Amazon Cloudfront produces access log that show all data requested by user. We will run a Hive Job which will analyze the data

1. Process your sample data by running a hive script
   1. Make your cluster is created completely and is in waiting state
   2. Add a new step with Step type as “Hive Program”, Name – “Process logs”, Script S3 location – “s3://us-west-2.elasticmapreduce.samples/cloudfront/code/Hive\_CloudFront.q”, Input S3 location – “s3://us-west-2.elasticmapreduce.samples”, output S3 bucket – “Hadoop-87” S3 bucket created earlier, Argument - “-hiveconf hive.support.sql11.reserved.keywords=false”
   3. Click Add to confirm and add

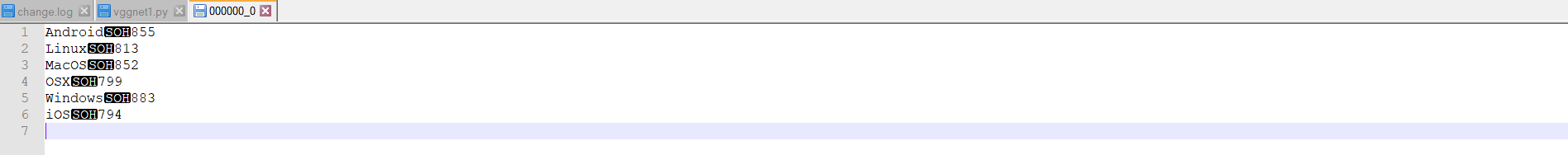
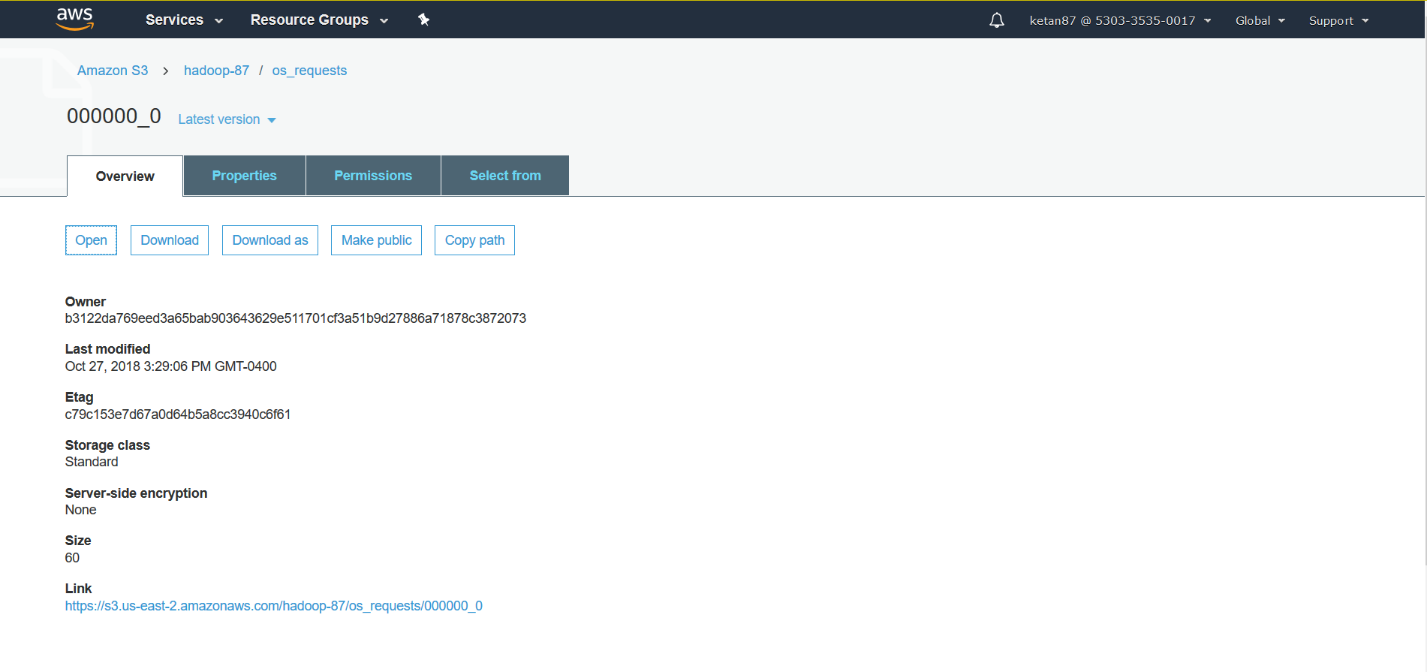




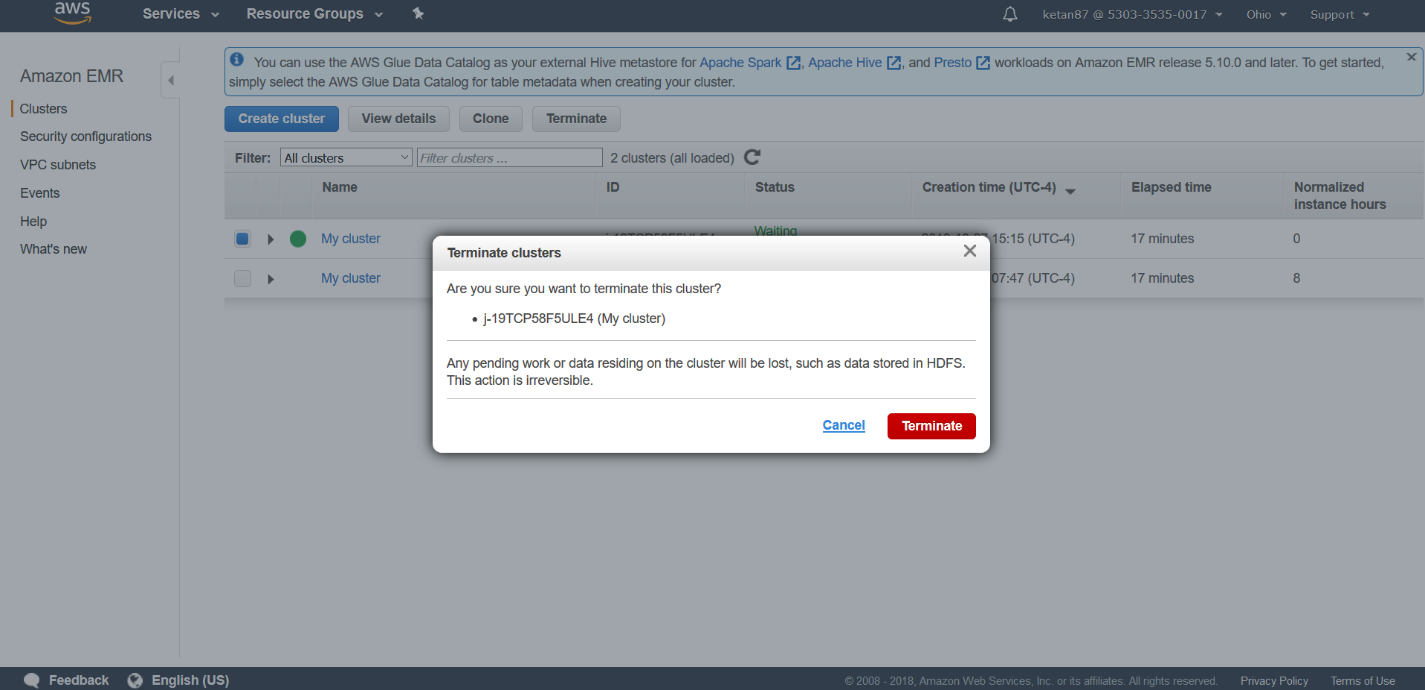
1. The Hive script submits a HiveQL query to retrieve the total per Operating system request for a given time frame which in this case is BETWEEN '2014-07-05' AND '2014-08-05'
2. Wait for the status to complete

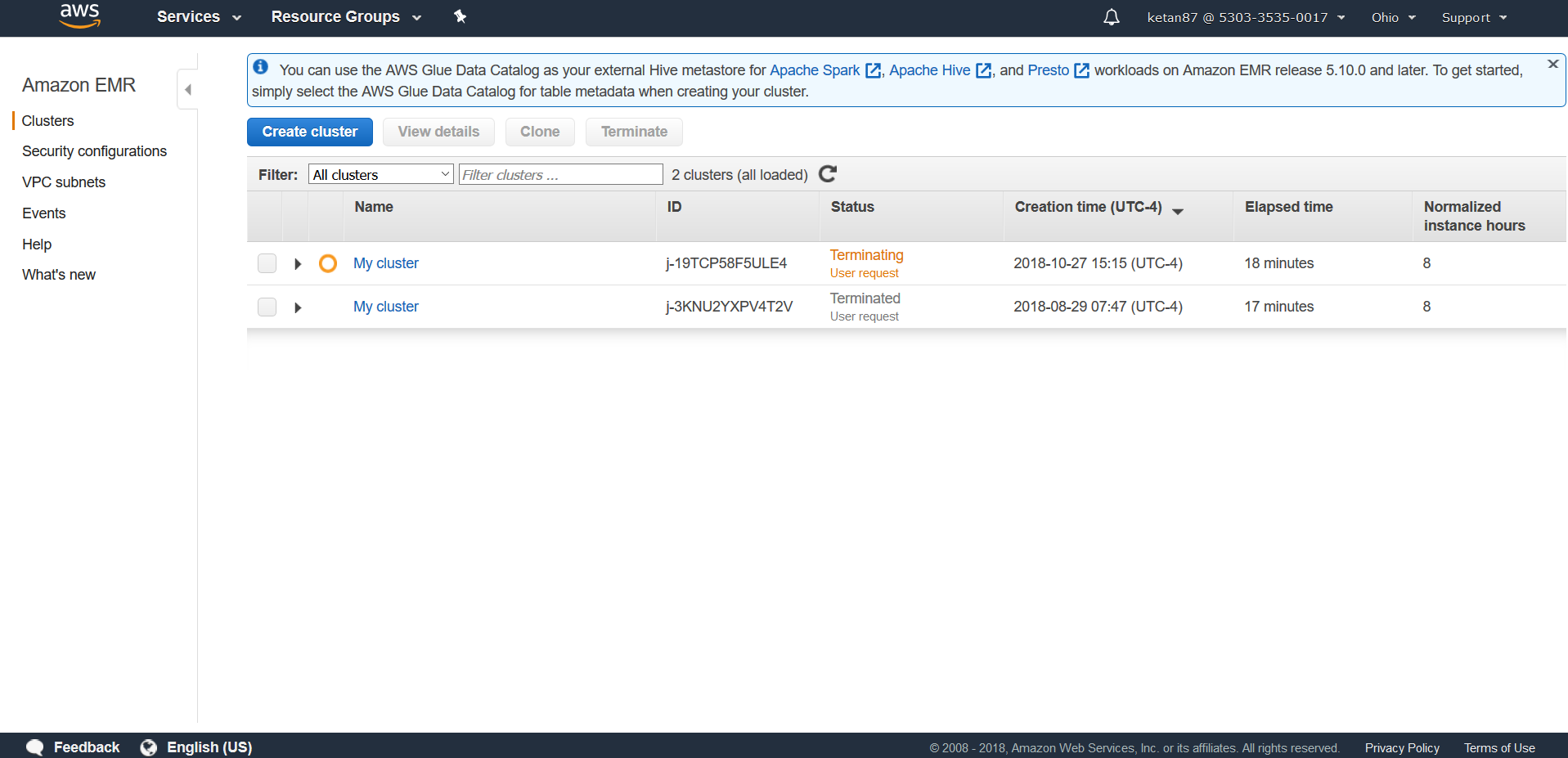


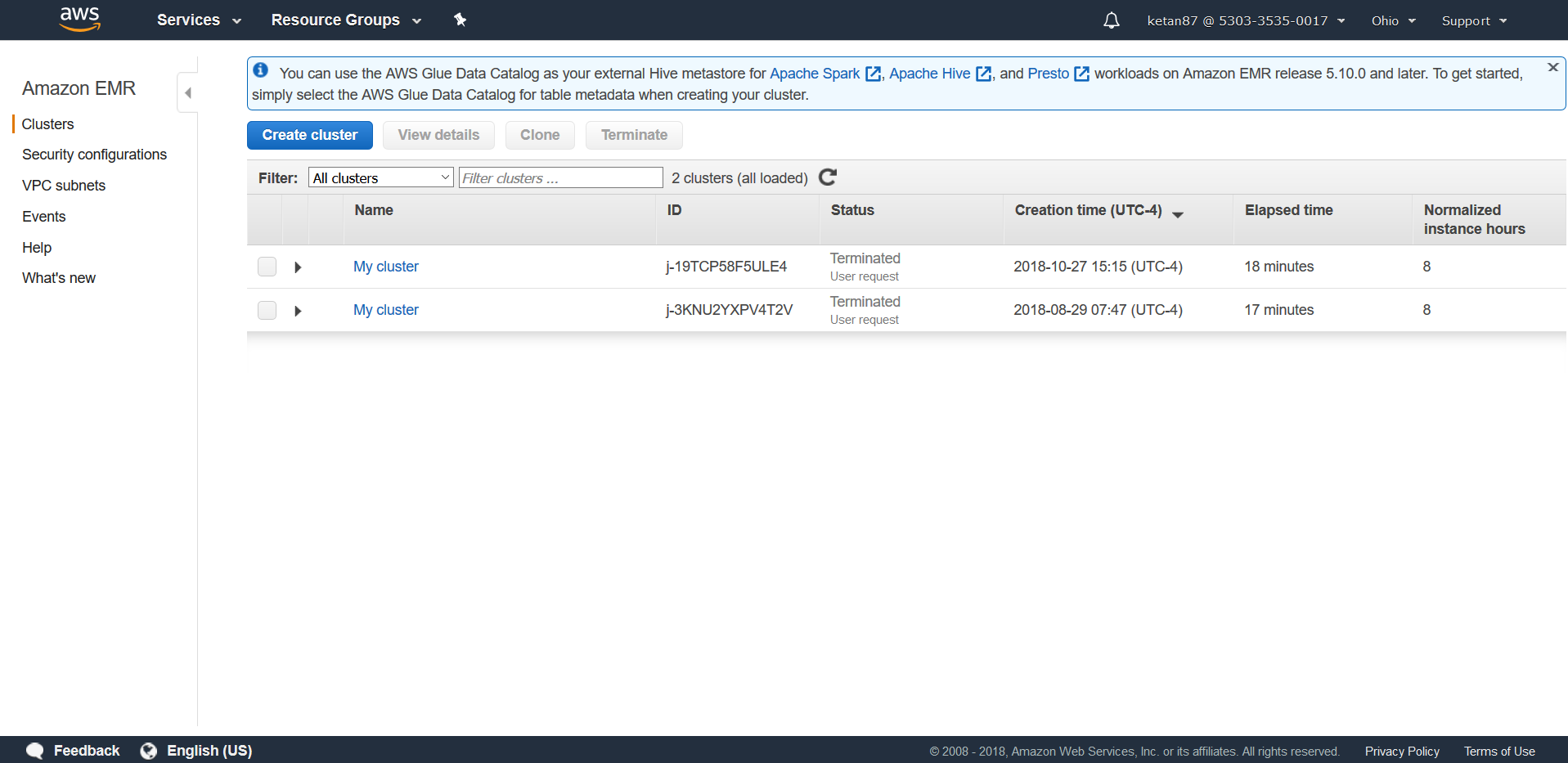
1. View the results.
   1. Open the Hadoop-87 bucket created earliest and open the “os\_requests” folder inside it.
   2. Select and download the “000000\_0” file
   3. Open the file in the text editor



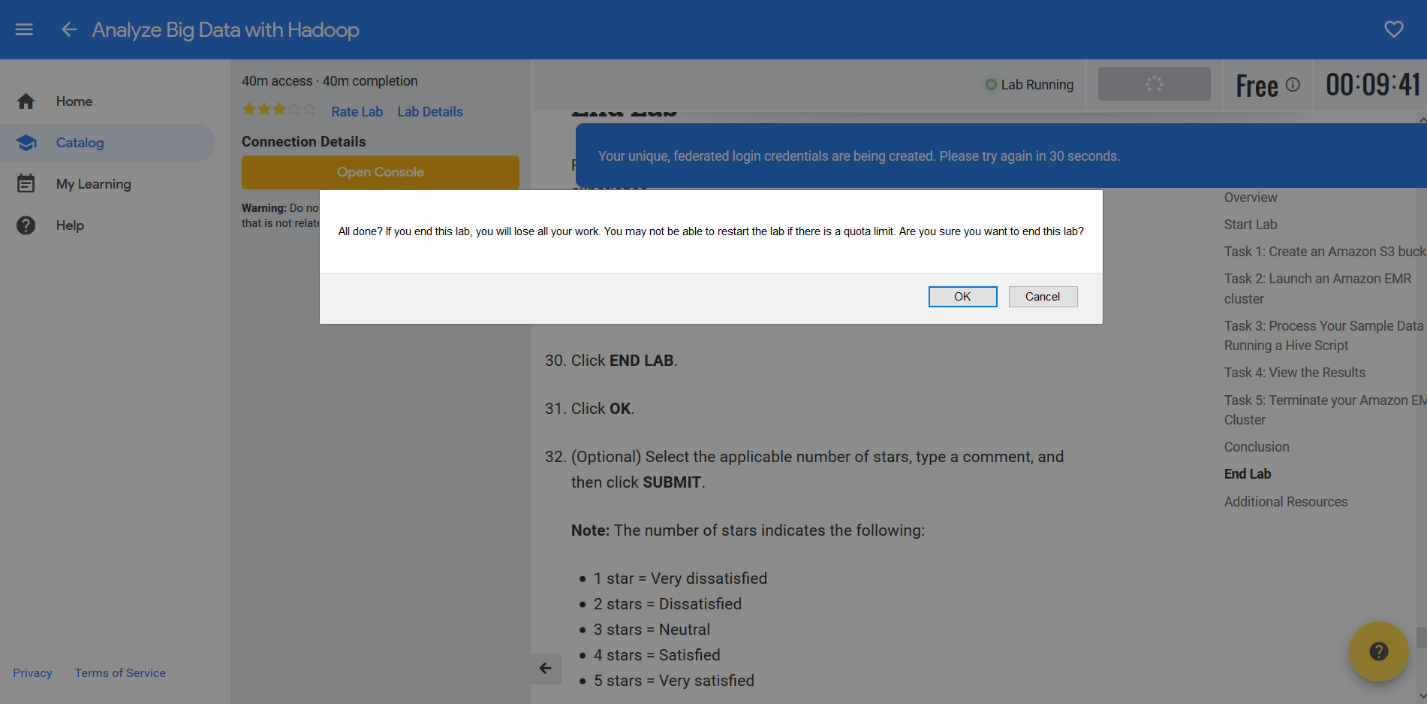
1. Terminate your Amazon EMR clusters and sign out of the AWS Console







1. Conclusion
   1. Launched a fully functional EMR Hadoop cluster
   2. Defined a schema and created a table for sample log data stored in S3
   3. Analyzed the data using HiveQL script and write back result in S3 bucket
   4. Download and view the result
2. Close the Lab



Total Time taken : 1 Hour