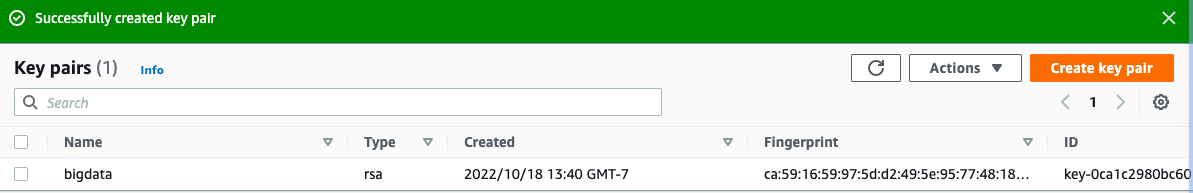
**HW 5 Data & Analytics**

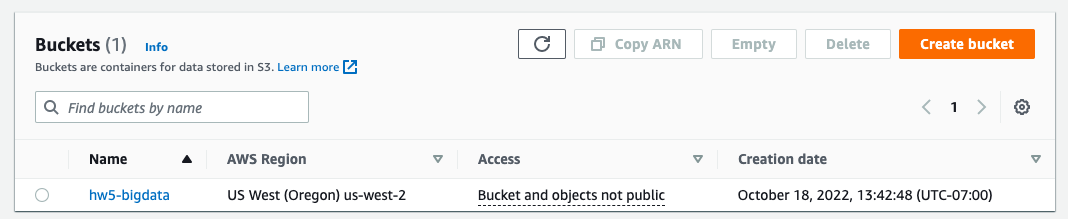
**Step 1: Plan and configure an Amazon EMR cluster**

In Amazon EC2 we are creating a key pair named big data.



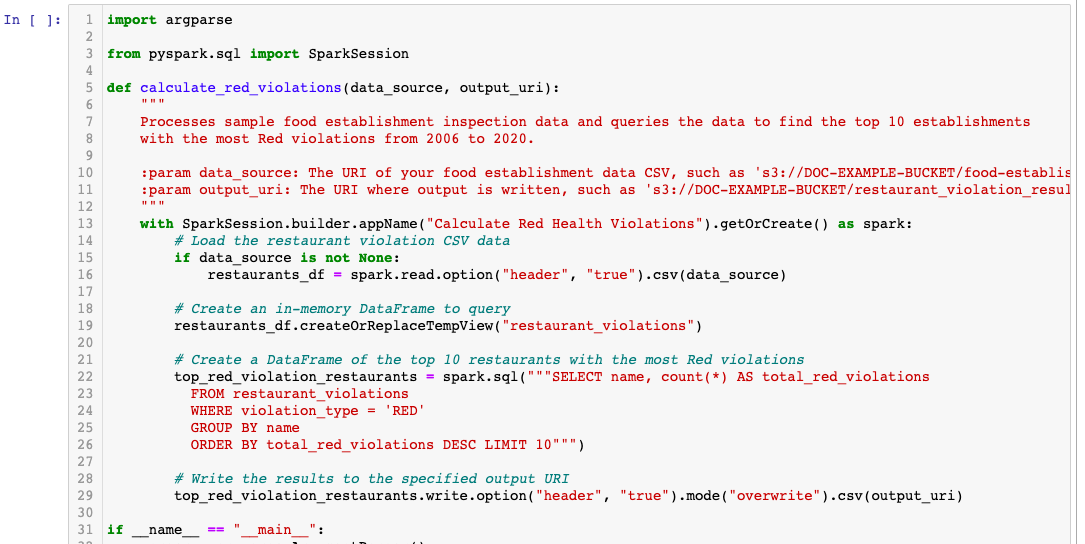
**Step 2: Creating Bucket**

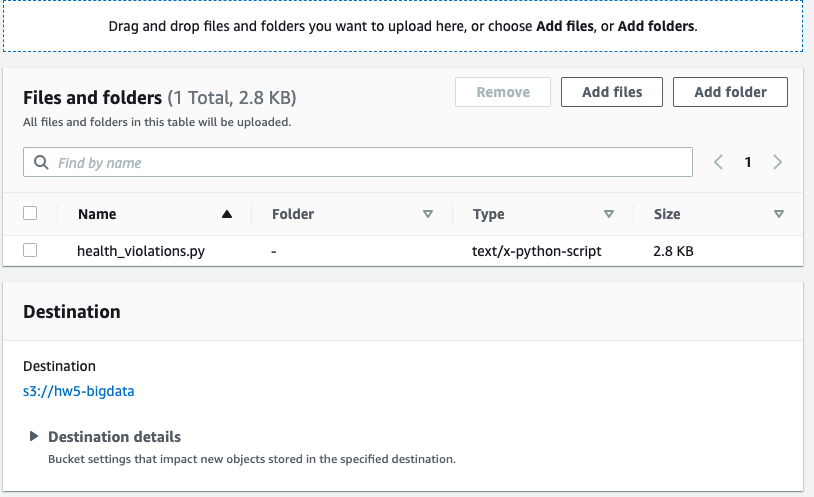
Creating a bucket named hw5-bigdata in the region us-west-2.



**Step 3: To prepare the example PySpark script for EMR**

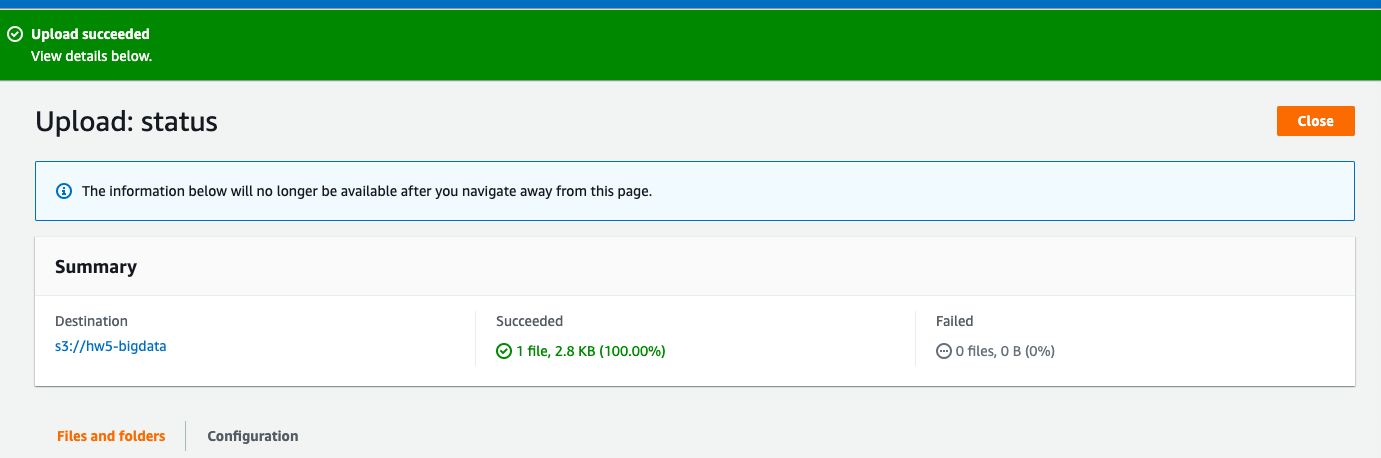
Copied and pasted the sample code into a new file and saved the file as health\_violations.py.





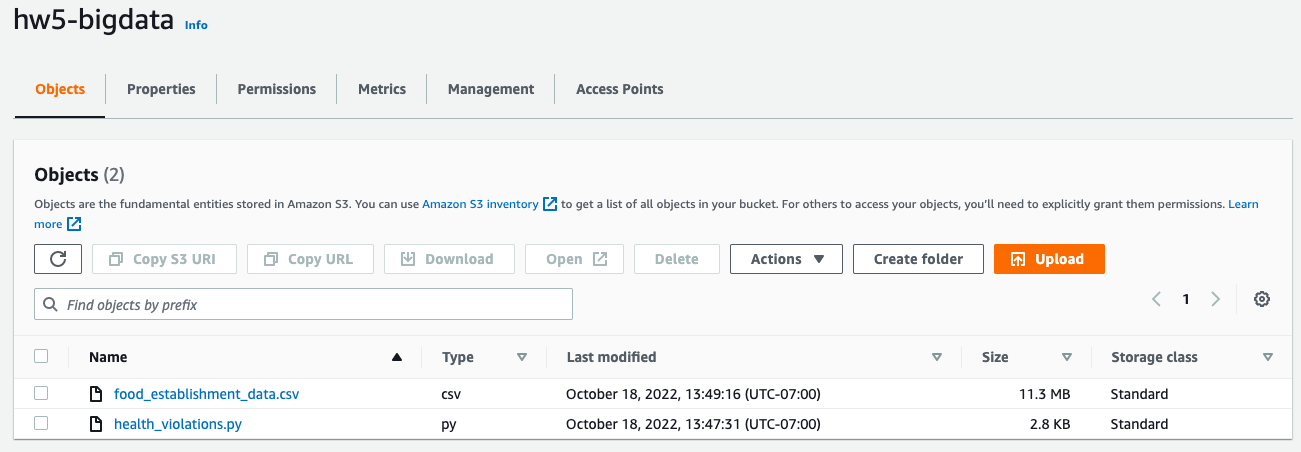
**Step 4: Upload Status**

Uploaded health\_violations.py to Amazon S3 into the bucket we created.



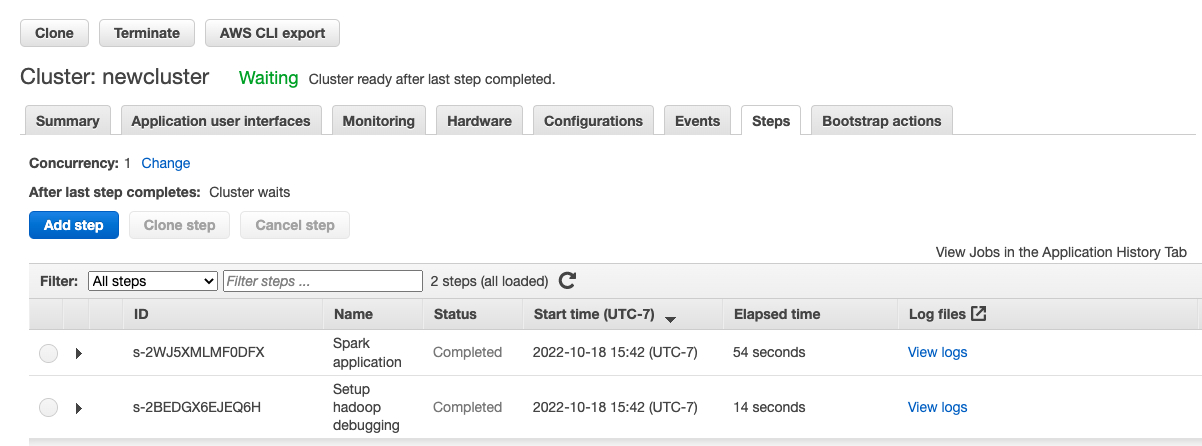
**Step 5: To prepare the sample input data for EMR**

Downloaded the zip file food\_establishment\_data.zip. Unzipped and saved food\_establishment\_data.zip as food\_establishment\_data.csv.



### **Step 6: Launch an Amazon EMR cluster**

### Firstly, we signed into the Amazon web service console using our team credentials and navigated to elasticmapreduce from the search bar. We choose to create a cluster and named the cluster as new cluster. Likewise, under the application option we choose spark.

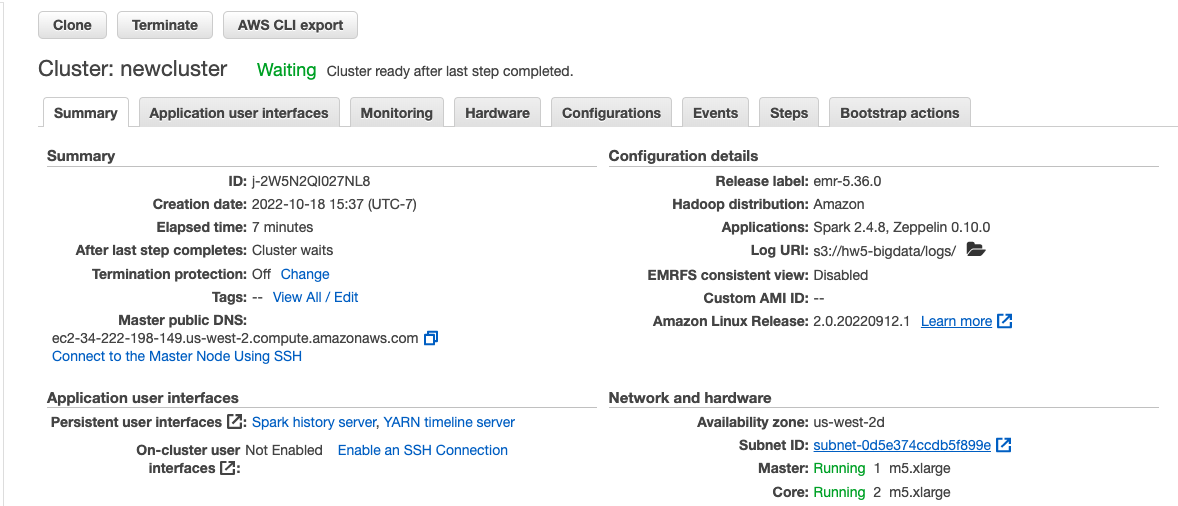


**Step 7: Manage your Amazon EMR cluster**

In the cluster list, we choose to add step under the Steps option. We selected spark application and selected default cluster in deploy node.

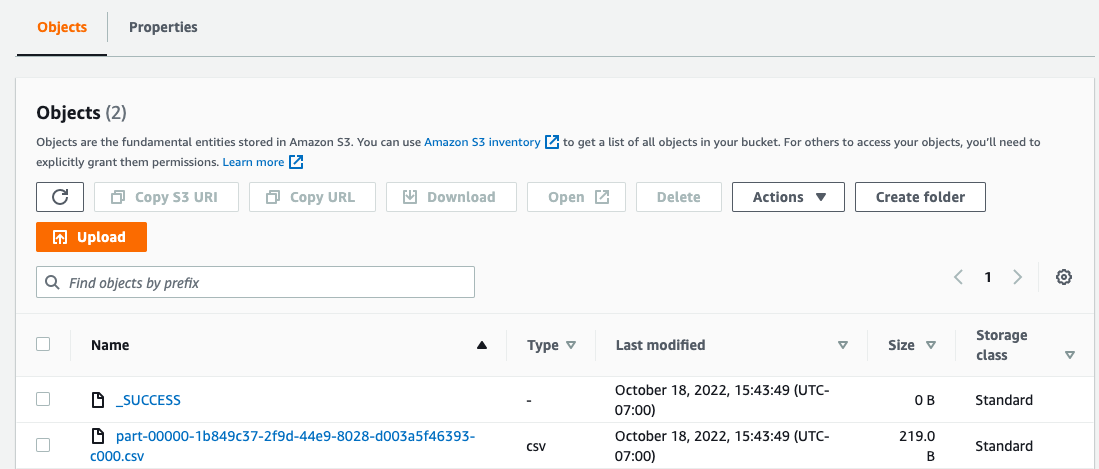
For the Application location we enter the location of our python file which was stored in amazon s3 bucket.

Status for the step changed from pending to complete after few minutes and it was ready.



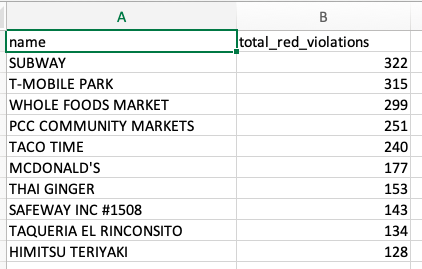
**Step 8: Verify**

The \_SUCCESS file verified that our operation was successful.



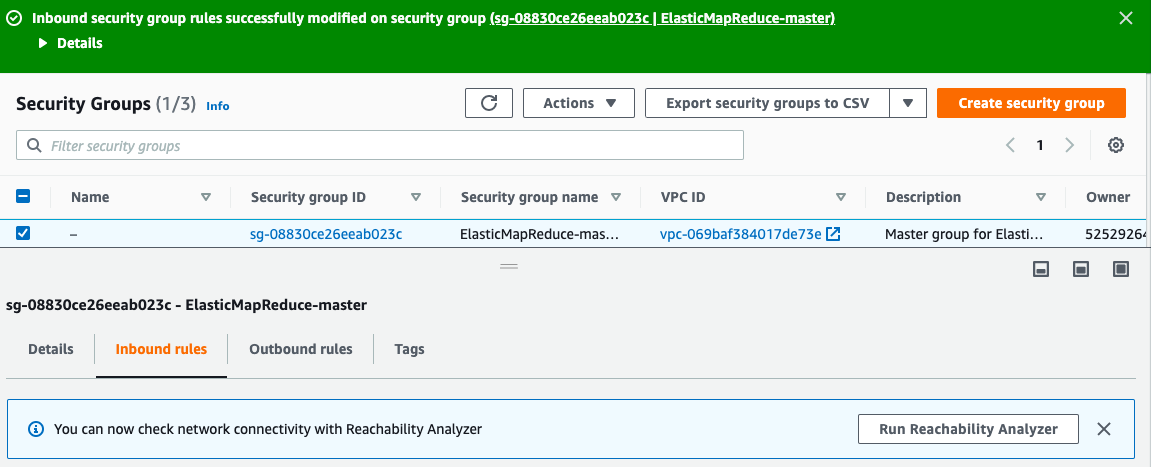
**Step 9: View result**

We navigated back to the Amazon S3 bucket to see the results and downloaded the health\_violations.py file.

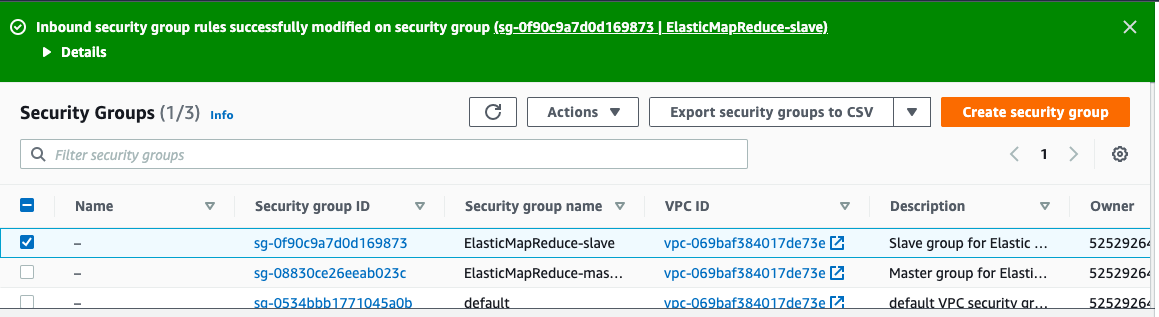


**Step 10: Authorize SSH connections**

To allow SSH access for trusted sources for the ElasticMapReduce-master security group with the old console.



To allow SSH access for trusted sources for the ElasticMapReduce-slave security group with the old console.



**Step 11: Clean up Resources**

Firstly, we navigated to the elasticmapreduce and selected the cluster to delete it. It took approximately 10 minutes to completely delete the cluster. A

