**Amazon EMR Lab**

**Mission: Launch an Amazon EMR cluster using the AWS Management console, and run the word count application to process data**

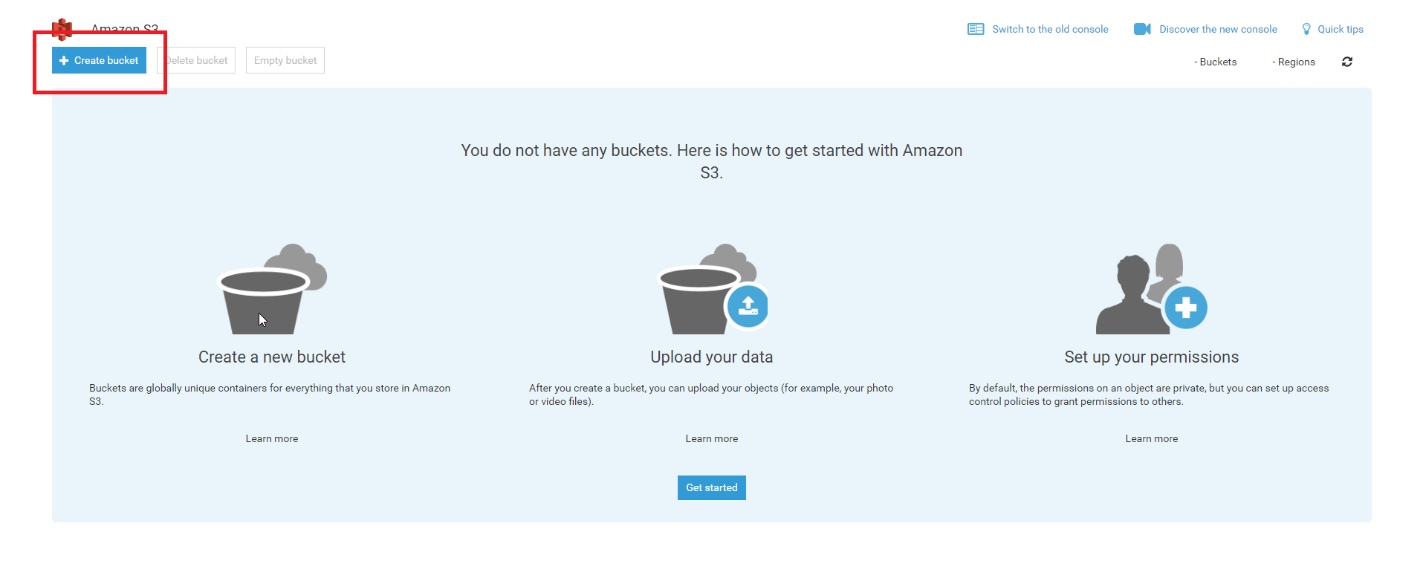
**Step 1: Create an Amazon S3 Bucket**

1. In the AWS Management console, click **Services** then click **S3**.

2. Click **Create Bucket**

**Note**

You are not charged for creating a bucket; you are charged only for storing objects in the bucket and for transferring objects in and out of the bucket. The charges you will incur through following the examples in this guide are minimal (less than $1).



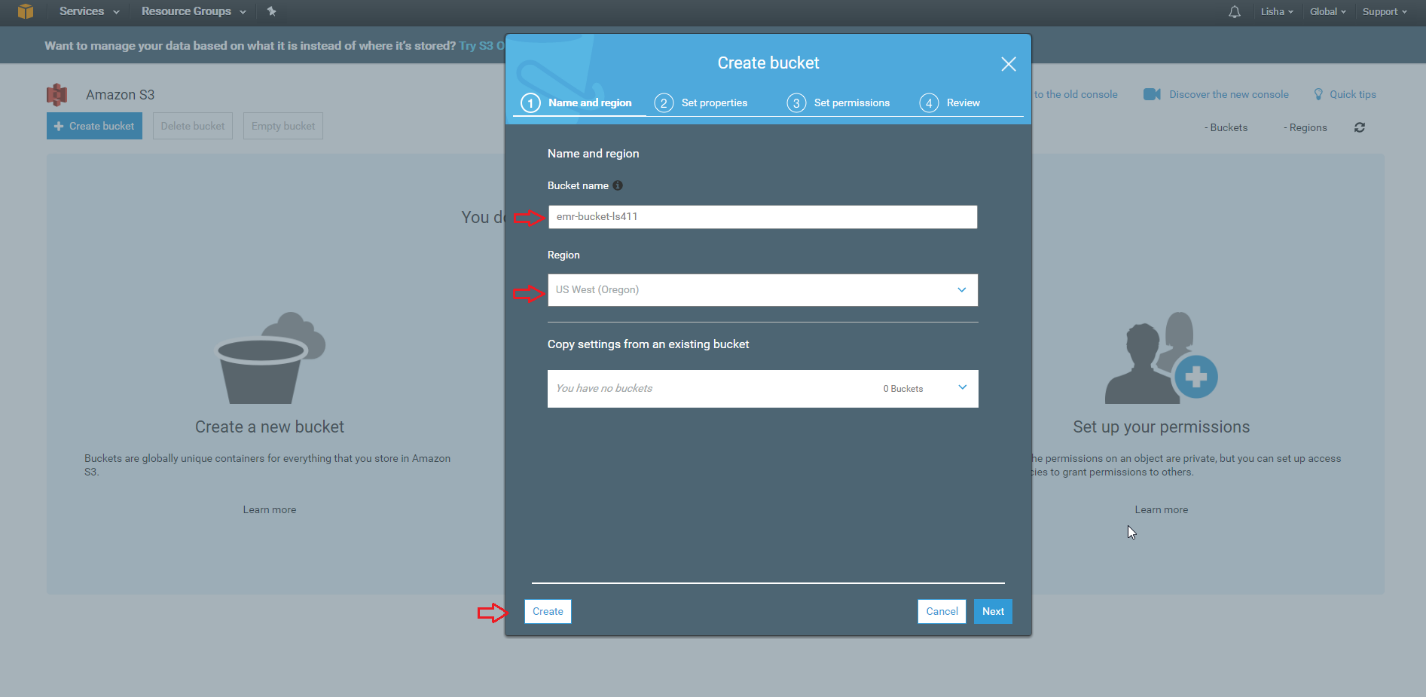
3. In the **Bucket Name** field, type **emr-bucket** followed by your initials and the date. For example, emr-bucket-ls411

**Tip**

* The name must be unique across all existing bucket names in Amazon S3.
* After you create the bucket you cannot change the name, so choose wisely.
* Choose a bucket name that reflects the objects in the bucket because the bucket name is visible in the URL that points to the objects that you're going to put in your bucket.

4. For **Region**, choose US West (Oregon) as the region where you want the bucket to reside.

5. Choose **Create.**

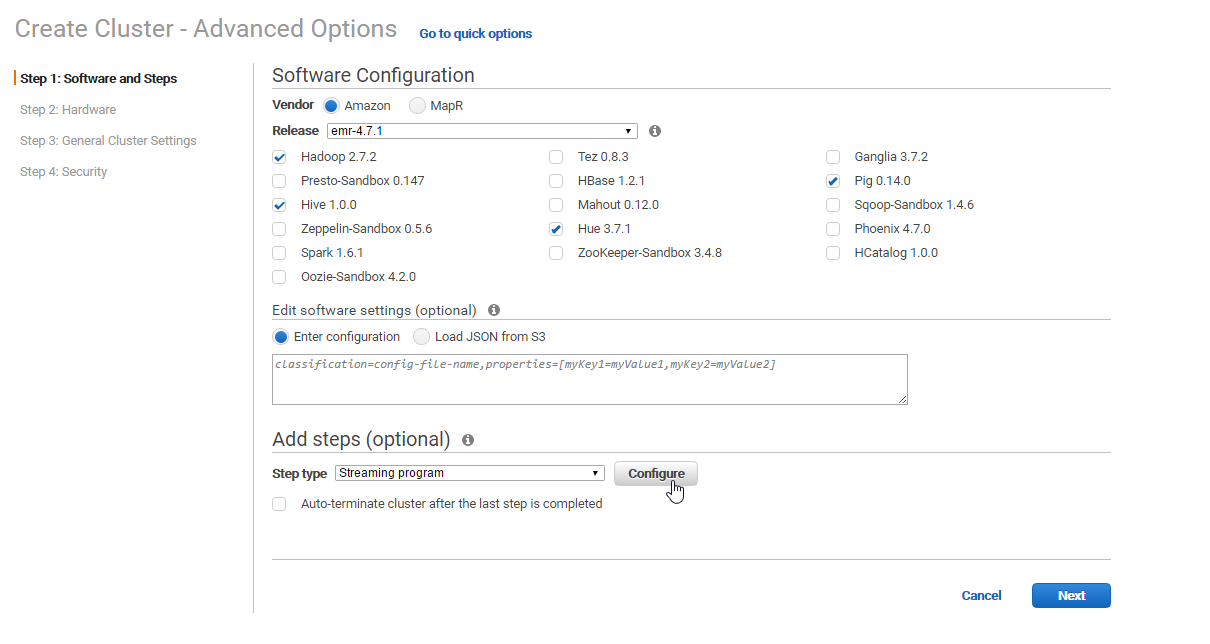


**Step 2: Creating and Launching a MapReduce Cluster**

6. In the AWS Management console, click **Services** then click **EMR**.

7. Click Create cluster.

8. Click **Go to advanced options** and for **Release**, click **emr-4.7.1**



9. On the Advanced Options page, under **Add Steps** (optional), for **Step type**, click **Streaming program**

10. Click **Configure**

11. For **Name**, type **Word count**

12. In the **Mapper** field, enter: **s3://elasticmapreduce/samples/wordcount/wordSplitter.py**

13. In the **Reducer** field, enter: **aggregate**

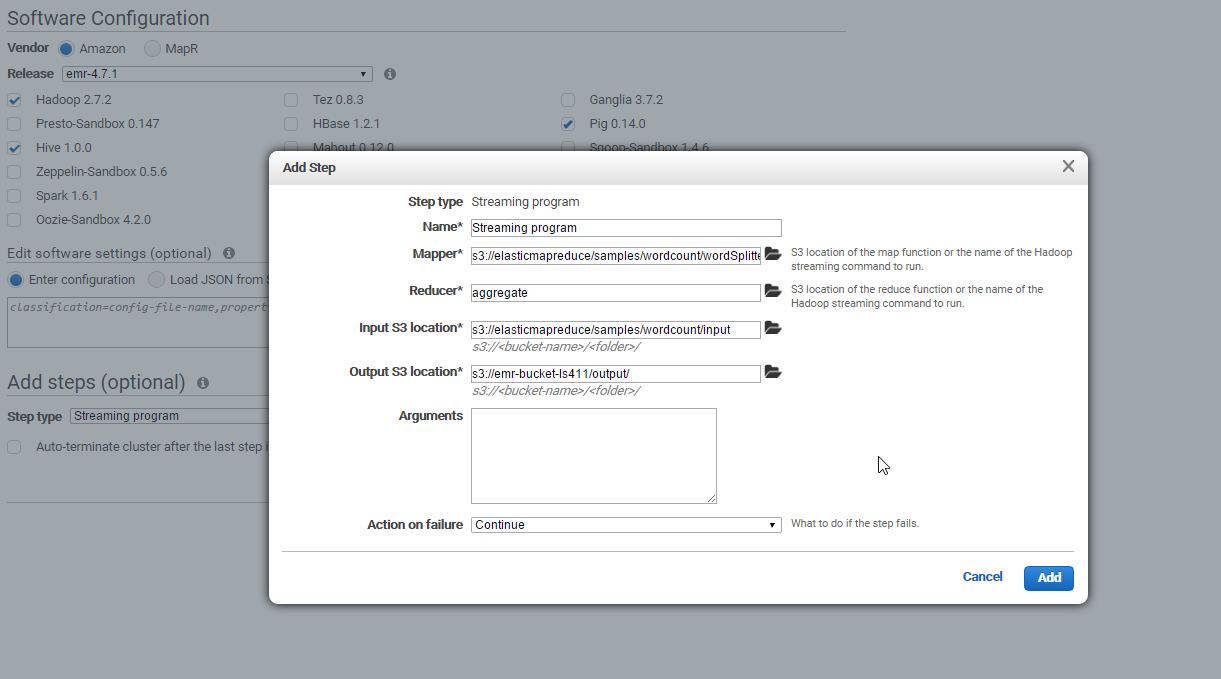
14. In the **Input S3 location** field, enter: **s3://elasticmapreduce/samples/wordcount/input**

For details and source code for the wordSplitter python script, see

<https://aws.amazon.com/articles/Elastic-MapReduce/2273>

15. In the Output S3 location field, type: **s3://<bucket-name>/output /** replacing <bucket-name> with the name of the Amazon S3 bucket you created earlier. Your URL should look similar to s3://emr-bucket-ls411/output/.

16. Click **Add**.



17. Check the box to select **Auto-terminate cluster after the last step is completed**.

18. Leave the hardware settings at their defaults. Click **Next**.

19. Under General Options, leave **Logging** enabled.

20. For the **S3 folder**, enter: **s3://<bucket-name>/logs/** replacing <bucket-name> with the name of the Amazon S3 bucket you created earlier. Your URL should look similar to s3://emr-bucket-ls411/logs/.

21. Click **Next**.

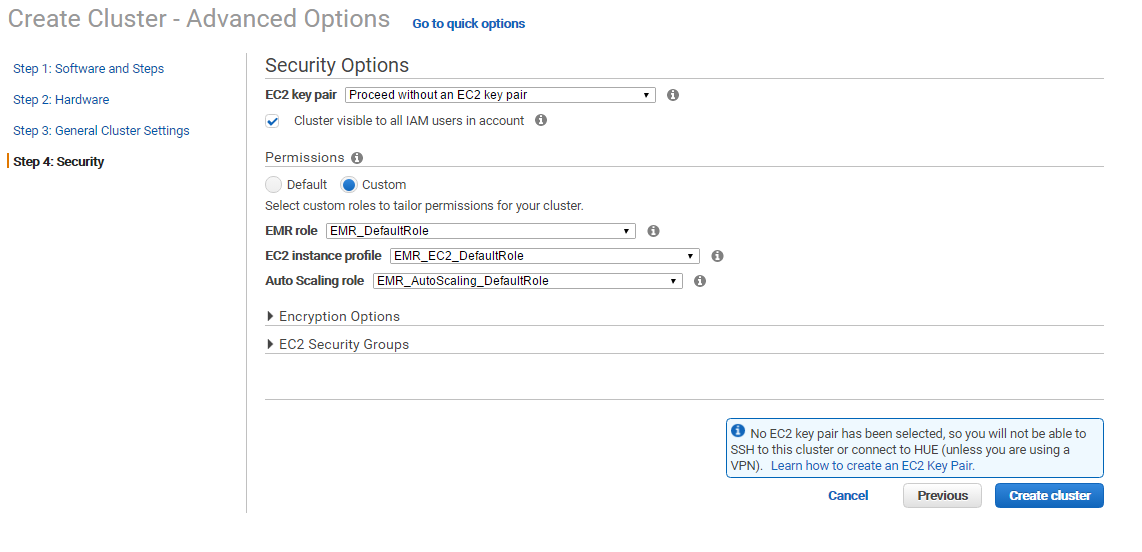
22. On the Security Options Page, in the **EC2 Key pair** field we will leave **Proceed without an EC2 key** pair. It is usually a best practice to choose an Amazon Ec2 key pair from the list. If you do not enter a value in this field, you cannot use SSH to connect to the master node.

23. For **Permissions**, click **Custom**.

24. For **EMR role**, click **EMRDefaultRole**.

25. For **EC2 instance profile**, click **EMREC2DefaultRole**.

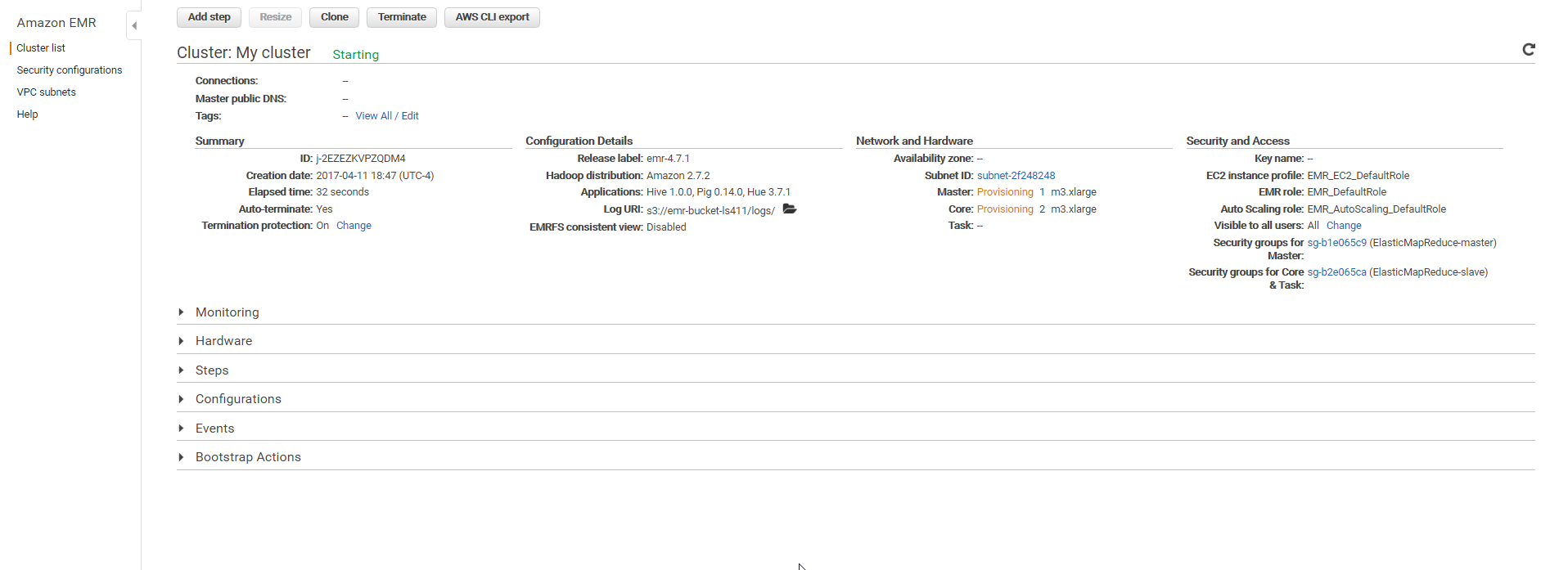
26. Leave all other settings at their defaults and click **Create cluster**.



**Step 3: Monitor the Cluster**

27. Click **Cluster List**. This shows a list of clusters and their status.

28. Click the arrow to the left of **cluster name** (“My cluster”) to see more details.



**Step 4: View the Results**

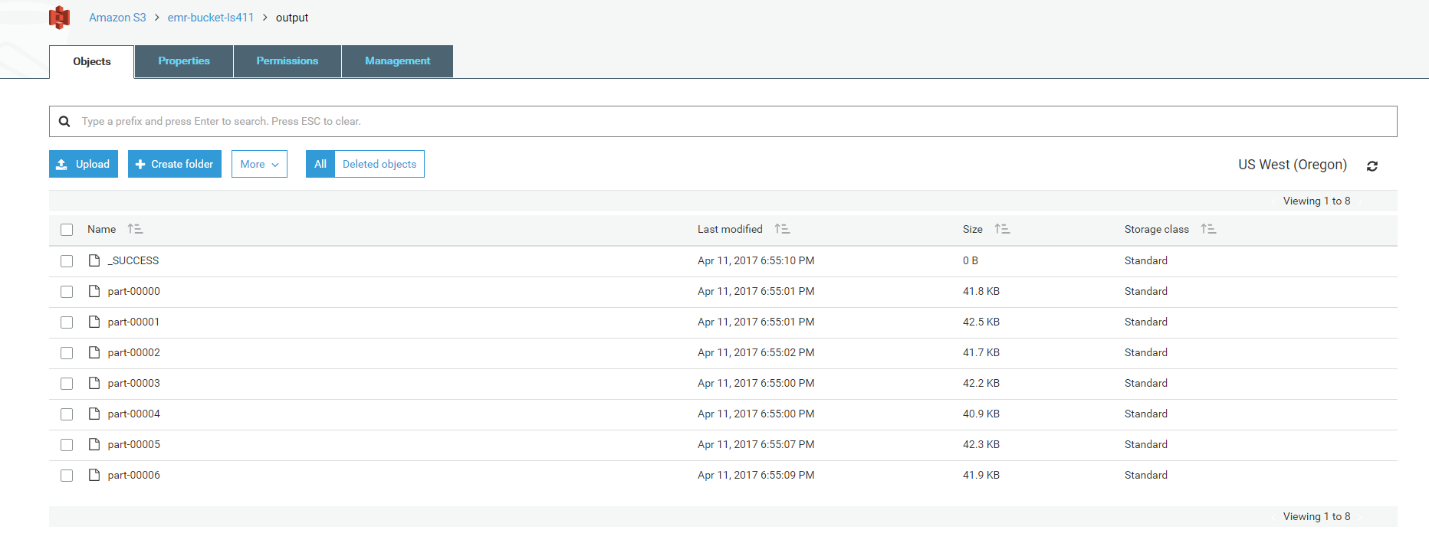
29. To view the results, go to the AWS Management Console. On the **Services** menu, click **S3**.

30. Select the bucket you created.

31. After selecting the bucket, click the **Output** folder.

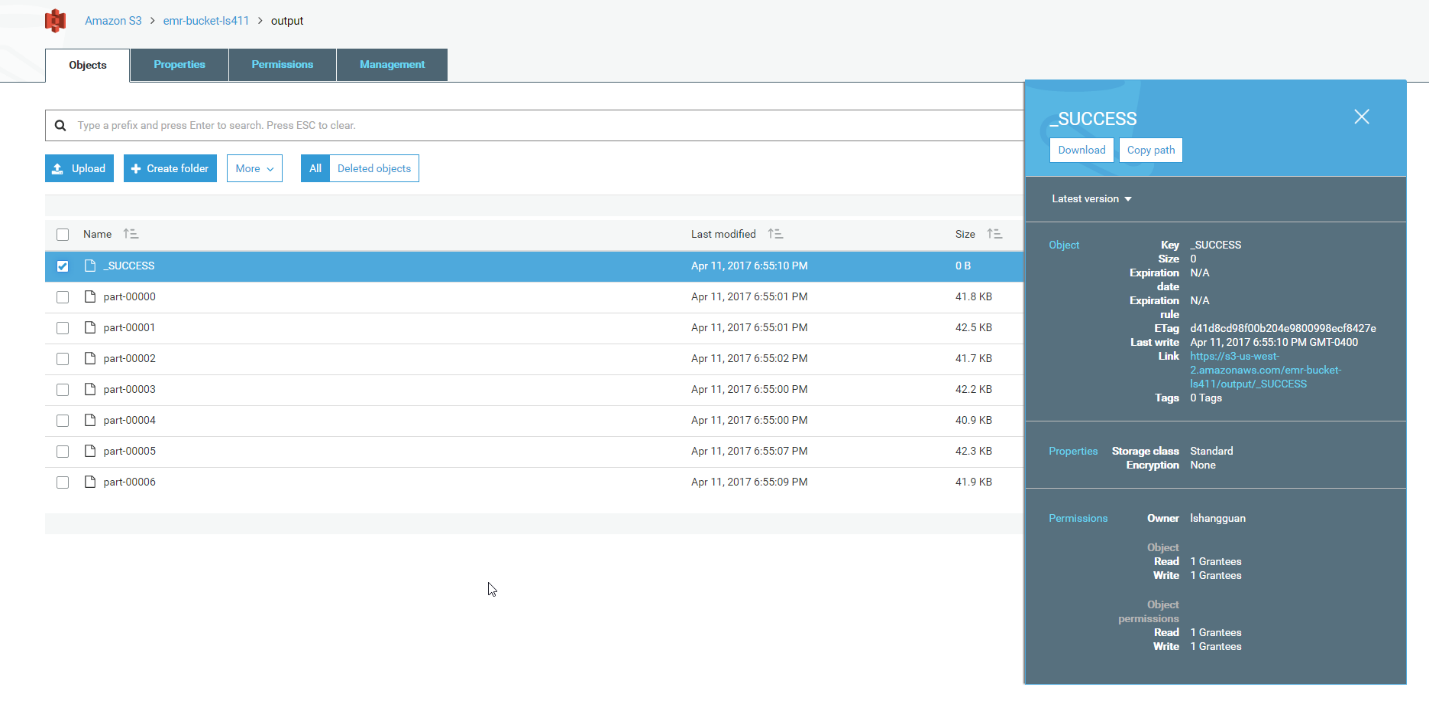
**Note**

The results of running the cluster are stored in test files. The first file in the listing is an empty file titled according to the result of the cluster. In this case, it is titled “\_SUCCESS” to indicate that cluster succeeded.



32. In the Output folder, you should see one or more files.

33. To download each file, right-click it, and select **Download**.



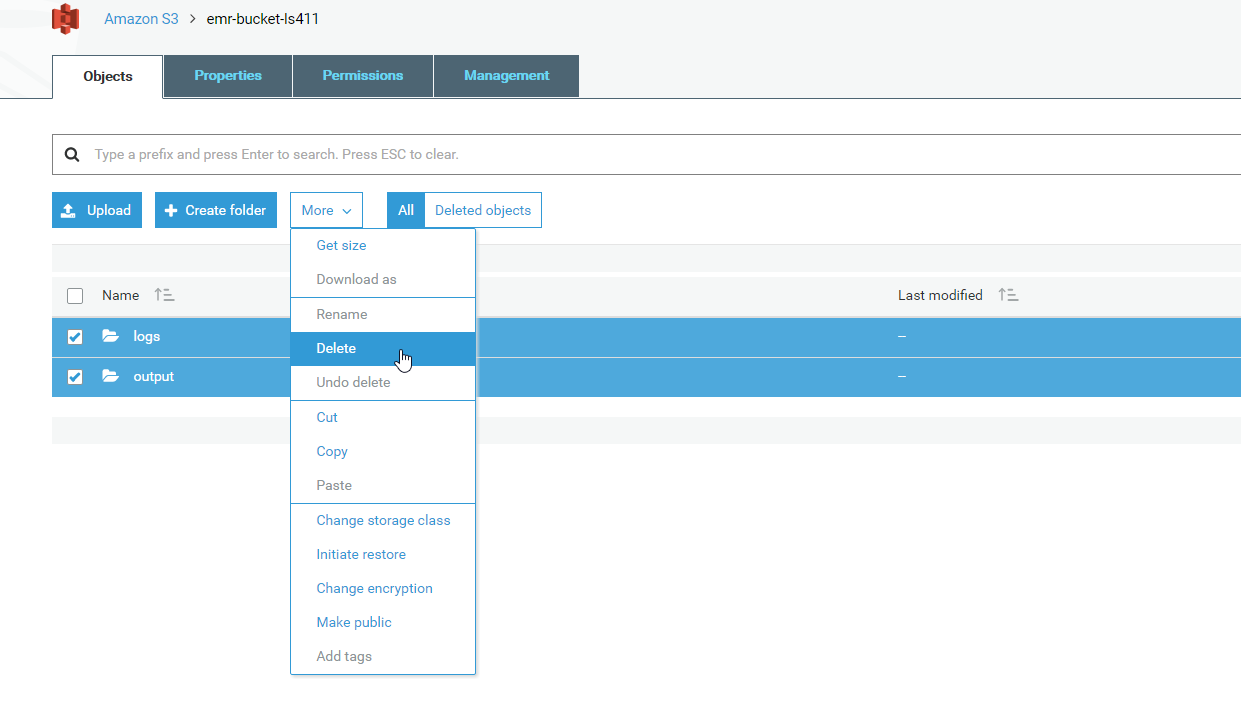
**Step 5: Reset Your Environment**

After you have completed this tutorial, you should remove your Amazon S3 bucket and terminate your Amazon EMR cluster to avoid incurring additional charge.

34. In the AWS Management console, click **Services** then click **S3**.

35. In the **Bucket name** list, choose the name of the bucket that you want to delete an object from.

36. In the **Name** list, select the check box next to the object that you want to delete, choose **More**, and then choose **Delete**.



36. In the **Delete objects** dialog box, verify that the name of the object you selected for deletion is listed, and then choose **Delete**.

37. In the **Bucket name** list, choose the bucket icon next to the name of the bucket that you want to delete and then choose **Delete bucket**.

38. In the **Delete bucket** dialog box, type the name of the bucket for delete confirmation and then choose **Confirm**.

