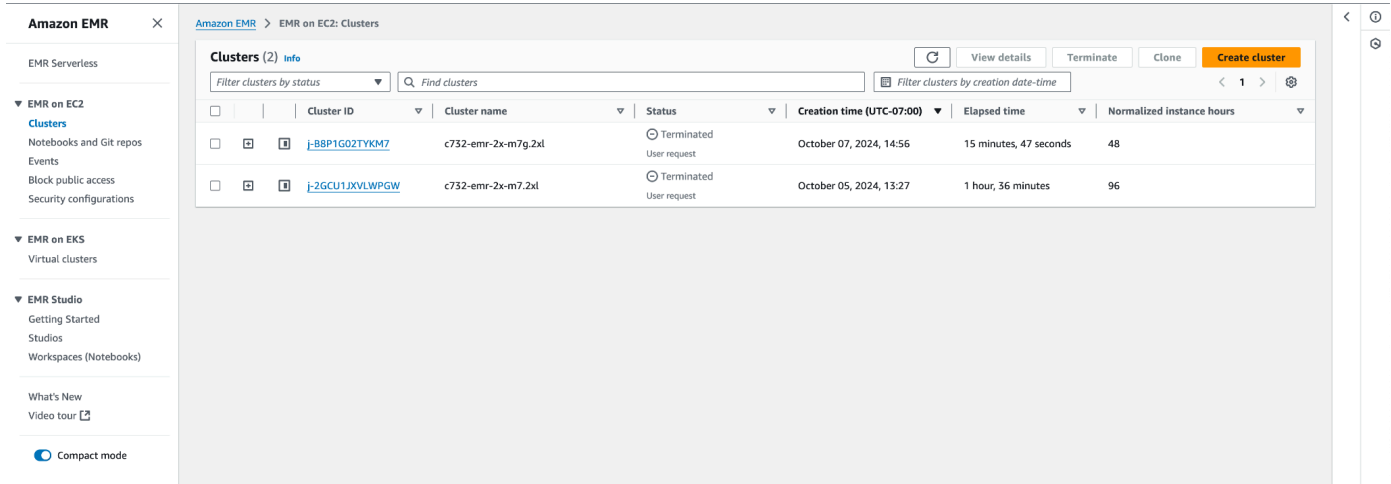


## Assignment 5 Answers

Take a screenshot of your list of EMR clusters (if more than one page, only the page with the most recent), showing that all have Terminated status.



The screenshot shows the Amazon EMR console interface. On the left is a navigation sidebar with options like 'EMR Serverless', 'EMR on EC2 Clusters', 'EMR on EKS', and 'EMR Studio'. The main panel displays a table of EMR clusters. The table has columns for Cluster ID, Cluster name, Status, Creation time, Elapsed time, and Normalized instance hours. Two clusters are listed, both with a status of 'Terminated User request'.

Cluster ID	Cluster name	Status	Creation time (UTC-07:00)	Elapsed time	Normalized instance hours
j-B8P1G0ZTYKM7	c732-emr-2x-m7g.2xl	Terminated User request	October 07, 2024, 14:56	15 minutes, 47 seconds	48
j-2GCU1JXVLWPGW	c732-emr-2x-m7.2xl	Terminated User request	October 05, 2024, 13:27	1 hour, 36 minutes	96

### For Section 2:

a. What fraction of the total data size (weather-1) was read into Spark when you used the "but different" data set to calculate the same result?

About 12% of the weather-1 data was read into Spark when I used the weather-1-but-different data set to calculate the same result.

b. What is different in the "but different" data that allows less data to be transferred out of S3 (and thus less S3 charges)? [hint]

The "but different" data set applies the concept of Hive-style partitioning. It partitions the data by the "observation" column. Therefore, when Spark is told that the input is coming from s3, the optimizer pushes back filtering to the file-reading phase, allowing less data to be transferred out of s3. As a result, when we perform a future query such as filtering on the observation, the query will skip partitions that don't match that observation.

### For Section 3:

Look up the hourly costs of the m7gd.xlarge instance on the EC2 On-Demand Pricing page. Estimate the cost of processing a dataset ten times as large as reddit-5 using just those 4 instances.

Total Uptime for processing reddit-5: 2.2min

Total Uptime for processing 10x reddit-5: 22min

Hourly cost of 1 m7gd.xlarge instance: \$0.2136

Hourly cost of 4 m7gd.xlarge instances: \$0.8544

Estimated cost of processing 10x reddit-5: approx. ~ \$0.3133