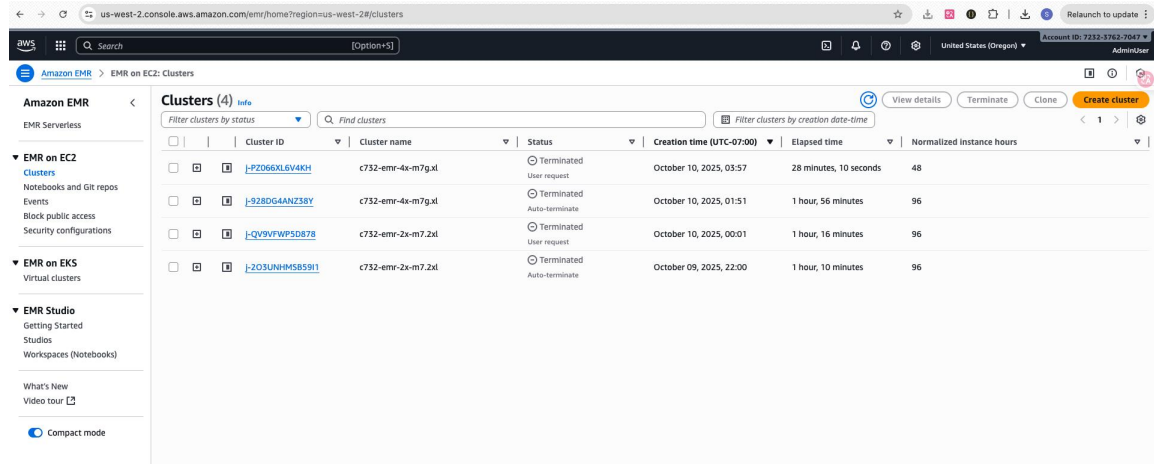


1. Take a screen shot of your list of EMR clusters (if more than one page, only the page with the most recent), showing that all have Terminated status.



Cluster ID	Cluster name	Status	Creation time (UTC-07:00)	Elapsed time	Normalized instance hours
<a href="#">j-p206xk1v4kh</a>	c732-emr-4x-m7g.xl	Terminated User request	October 10, 2025, 03:57	28 minutes, 10 seconds	48
<a href="#">j-9380g4anz38y</a>	c732-emr-4x-m7g.xl	Terminated Auto-terminate	October 10, 2025, 01:51	1 hour, 56 minutes	96
<a href="#">j-qv9vfwps0878</a>	c732-emr-2x-m7.2xl	Terminated User request	October 10, 2025, 00:01	1 hour, 16 minutes	96
<a href="#">j-z03unhms8591</a>	c732-emr-2x-m7.2xl	Terminated Auto-terminate	October 09, 2025, 22:00	1 hour, 10 minutes	96

2. 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?

- Input for weather-1: 294.0 KiB
- Input for weather-1-but-different: 26.4 KiB

$$26.4/294=0.0898$$

So approximately 9%

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

but-different datasets are partitioned by directory on S3 (for example, a separate structure like observation=ADPT/...).

Spark/EMR performs partitioning (partition pruning) and filter pushdown when reading: only the partition directories relevant to the query are read, skipping irrelevant directories and files.

This results in fewer objects being pulled from S3, fewer bytes being transferred, and naturally smaller input files, resulting in higher S3 costs.

3. 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. Since my EMR step did not complete successfully, the runtime estimate is based on my local-cluster execution.

On my local Spark environment (4 executors × 1 core each), the job relative\_score\_bcast.py finished processing the full reddit-5 dataset in about 6 minutes (0.1 hours).

For the cost estimation, we assume running on four m7gd.xlarge instances, each with 4 vCPUs and 16 GiB RAM.

According to the AWS EC2 On-Demand Pricing page, an m7gd.xlarge instance costs \$0.2136 per hour (in us-west-2).

#### Scaling estimate

- Dataset size:  $10 \times$  larger
- Compute capacity:  $\approx 4 \times$  larger
- $\rightarrow$  Expected runtime  $\approx (10 / 4) \times 6 \text{ min} \approx 15 \text{ minutes (0.25 hours)}$

#### Cost calculation

$$4 * \$0.2136 * 0.25 = \$0.21$$