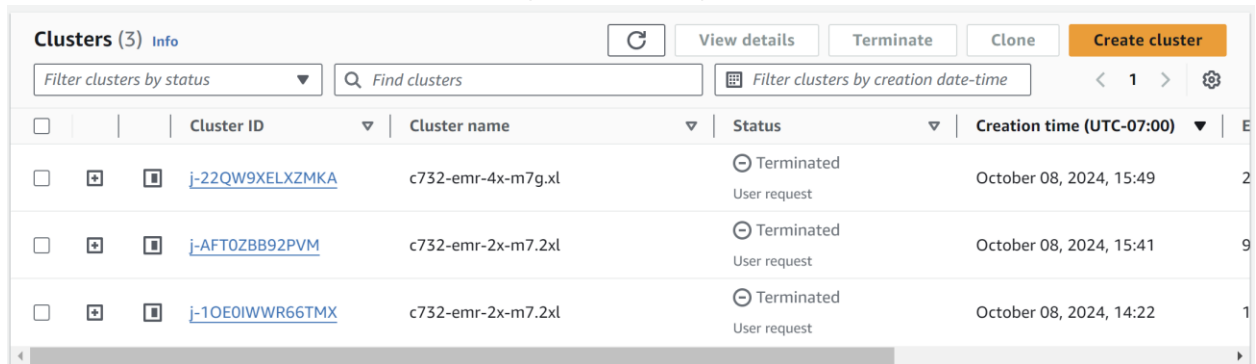


1. The middle cluster was created incorrectly, which is why I have 3 clusters.



The screenshot shows the AWS EMR Clusters console. At the top, there are buttons for 'View details', 'Terminate', 'Clone', and 'Create cluster'. Below these are filters for 'Filter clusters by status' and 'Find clusters'. The main table lists three clusters, all with a status of 'Terminated' and 'User request' as the reason. The clusters are:

Cluster ID	Cluster name	Status	Creation time (UTC-07:00)
j-22QW9XELXZMKA	c732-emr-4x-m7g.xl	Terminated User request	October 08, 2024, 15:49
j-AFT0ZBB92PVM	c732-emr-2x-m7.2xl	Terminated User request	October 08, 2024, 15:41
j-10E0IWWR66TMX	c732-emr-2x-m7.2xl	Terminated User request	October 08, 2024, 14:22

2. a)
Weather-1 input size: 2.6MiB
Weather-but-different input size: 327.6KiB
 $327.6\text{KiB} / 2.6\text{MiB} = 123\text{KiB}$
b)
The different file uses hive-partitioning. Hive-partitioning separates data into folders based on partition keys. When querying data, the system can skip over irrelevant partitions and only read the relevant ones, reducing the amount of data retrieved.
3. Total uptime for reddit-5 on cluster: 13min (3.9, 2.7, 2.8, 3.6mins per job)
Uptime on AWS for reddit-5: 2.8 min
Time to run a dataset 10x as large = 28min
M7gd.xlarge cost: 0.2136/hr
Cost = $(28\text{min} / 60\text{min} * 4 \text{ instances}) * 0.2136/\text{hr} = \sim 40 \text{ cents}$