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.

A screenshot of a computer

Description automatically generated

1. For Section 2:
   1. What fraction of the input file was prefiltered by S3 before it was sent to Spark?

The original run showed the following data:

Input Size / Records: 2.6 MiB / 3245

Output Size / Records: 27.2 KiB / 3245

The run with S3 filtering showed the following data:  
  
Input Size / Records: 97.7 KiB / 3245

Output Size / Records: 27.2 KiB / 3245

If we assume that 2.6 \* 1000 = 2600 KiB is the original size of the input file, and with S3 filtering, the input is 97.7 KiB, then 1 – (97.7 KiB / 2600 KiB) = 0.9624, or approximately 96% of the data was prefiltered before it was sent to Spark. That’s pretty impressive.

* 1. Comparing the different input numbers for the regular version versus the prefiltered one, what operations were performed by S3 and which ones performed in Spark?

1. 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. If you wanted instead to process this larger dataset making full use of 16 instances, how would it have to be organized?

**Cost analysis:**

For the cost analysis, I will ignore start-up times and focus on the actual run-time of jobs.

The cost per instance-hour is $0.2136. We used four instances, and my Total Uptime was 4.9 minutes.

The portion of an hour elapsed is given by 4.9/60 = 0.0817.

The cost per instance is then $0.2136 \* 0.0817 = $0.01744.

The total cost for all four instances is given by $0.01744 \* 4 = $0.069776

Considering that the dataset was ten times larger, we might assume that the portion of an hour elapsed is 4.9 \* 10 = 49 minutes. So, the portion of an hour elapsed would be given by 49/60 = 0.817.

The cost per instance would be $0.2136 \* 0.817 = $0.1744.

And the total cost for all four instances would be $0.1744 \* 4 = $0.69776, or $0.70.

So, we could estimate the cost of processing a dataset ten times as large as reddit-5 using just those four instances at 70 cents.

**Full use of 16 instances analysis:**

Suppose we want to use 16 instances, each with four cores. In that case, we need to recognize that the degree of parallelism available to us is 16 instances \* 4 cores per instance = 64 parallelizable cores. We want the number of partitions to be between 1-2 orders of magnitude greater than the number of available cores. Hence, it would be appropriate for the input data to be split between 640 – 6400 separate files.