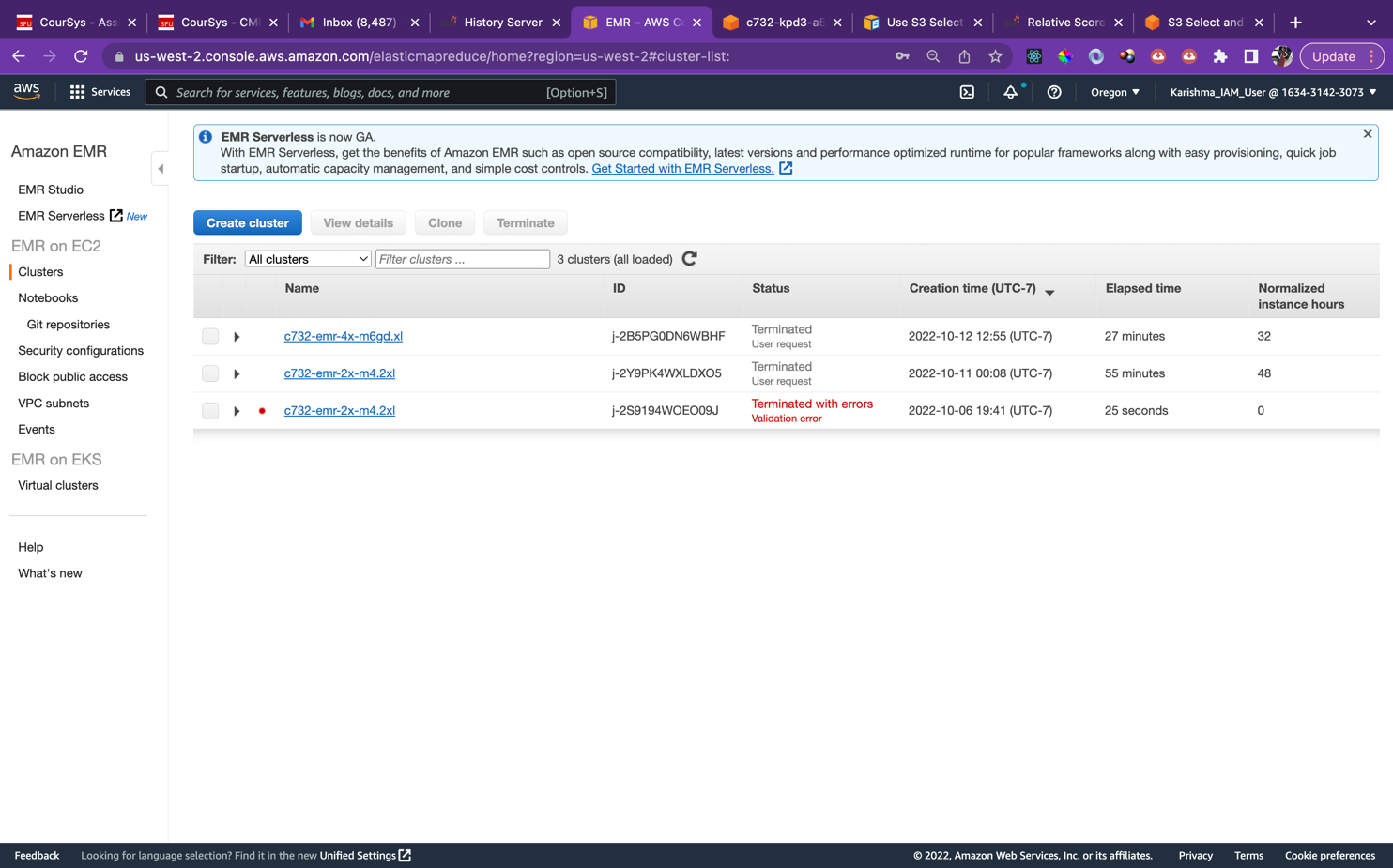
Q1) The terminated clusters are as shown below.



Q2) a) The input size without filtering = 2.6 MiB

The input size with S3 filtering = 97.7 KiB

The fraction of the input file was filtered and sent to Spark = 97.7/2662.4 = 0.0366

Thus the amount that was filtered out by S3Select = 96.34% and the amount that was sent to spark after filtering = 3.66%

Text

Description automatically generated

Without S3 Select

Text

Description automatically generated

With S3 Select

b) The SQL operations such as filtering, where and select, etc are being performed through S3Select whereas the other operations are performed by Spark.

Q3) From the Jobs and DAGs below we can see that the maximal amount of time is taken through the collect job and the runJob, both involve reading from the input and writing into the output, respectively. Since most time is spent in IO operations compared to the rest of the computations, we can say that the application is IO bound. Table

Description automatically generated

Diagram

Description automatically generated Diagram

Description automatically generated

Collect DAG runJob DAG

b)

cost of m6gd.xlarge = $0.1808/hour of usage while using 4vCPUs

Since the cost is calculated depending on the time taken, a dataset 10 times the size of reddit-5 would take much more time to execute with the same 4 instances and hence the price would increase accordingly.

In order to process a very large dataset while making use of 16 instances, we could partition the dataset so that the instances can work parallelly with these chunks thus making better use of the available resources.