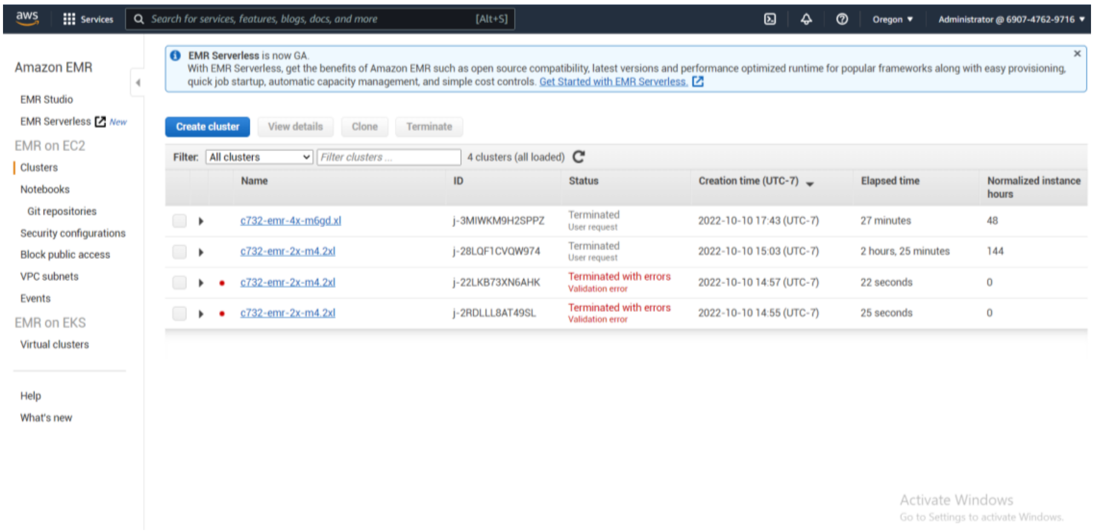
Use a word processor of your choice to generate answers.pdf and fill in the following:

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.



EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. Get Started with EMR Serverless.

Filter: All clusters Filter clusters ... 4 clusters (all loaded)

		Name	ID	Status	Creation time (UTC-7)	Elapsed time	Normalized instance hours
□	▸	c732-emr-4x-m6gd.xl	j-3MIWKM9H2SPPZ	Terminated User request	2022-10-10 17:43 (UTC-7)	27 minutes	48
□	▸	c732-emr-2x-m4.2xl	j-28LQF1CVQW974	Terminated User request	2022-10-10 15:03 (UTC-7)	2 hours, 25 minutes	144
□	▸ ●	c732-emr-2x-m4.2xl	j-22LKB73XN6AHK	Terminated with errors Validation error	2022-10-10 14:57 (UTC-7)	22 seconds	0
□	▸ ●	c732-emr-2x-m4.2xl	j-2RDLLL8AT49SL	Terminated with errors Validation error	2022-10-10 14:55 (UTC-7)	25 seconds	0

2. For Section 2:

   a. What fraction of the input file was prefiltered by S3 before it was sent to Spark?

      Ans:

      The input size with S3 prefiltering is 97.7 KiB

      The input size without S3 prefiltering is 2.6MiB

      Hence the fraction should be $1 - \frac{97.7}{2.6*1024}\% \approx 96.3\%$

   b. 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?

      Ans:

      (Cite: AmazonS3/latest/userguide)

      According to the user guide, S3 can only support SELECT SQL command. Hence in this case, S3 filtered several single columns according to the conditions given by Python. (.where() or .filter())

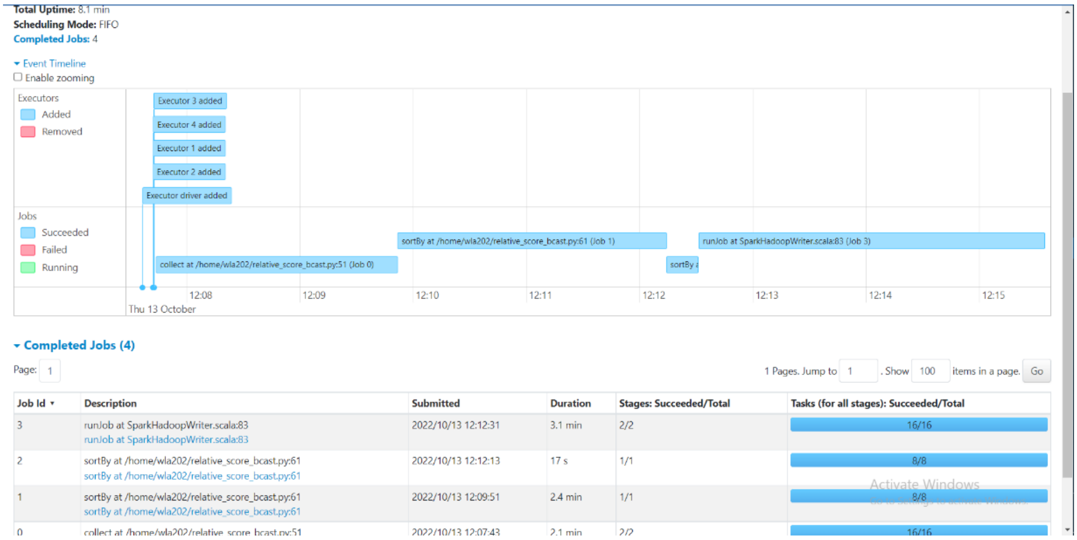
      Spark did calculation and added a new column to DataFrame(It's .withColumn() in my code.).

      Spark also used .select() for transformation, it returned a new DataFrame with three columns.

3. For Section 3:

   a. Reviewing the job times in the Spark history, which operations took the most time? Is the application IO-bound or compute-bound?

      Ans:



Total Uptime: 8.1 min
Scheduling Mode: FIFO
Completed Jobs: 4

▾ Event Timeline
□ Enable zooming

▾ **Completed Jobs (4)**

Page: 1 1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
3	runJob at SparkHadoopWriter.scala:83 runJob at SparkHadoopWriter.scala:83	2022/10/13 12:12:31	3.1 min	2/2	16/16
2	sortBy at /home/wla202/relative_score_bcast.py:61 sortBy at /home/wla202/relative_score_bcast.py:61	2022/10/13 12:12:13	17 s	1/1	8/8
1	sortBy at /home/wla202/relative_score_bcast.py:61 sortBy at /home/wla202/relative_score_bcast.py:61	2022/10/13 12:09:51	2.4 min	1/1	8/8
0	collect at /home/wla202/relative_score_bcast.py:51	2022/10/13 12:07:43	2.1 min	2/2	16/16

As we can see, three operations took the most time, they are ‘collect’, ‘sortBy’, and ‘runJob’.

‘collect’ and ‘runJob’ are IO operations and ‘sortBy’ is computation.

The total times of ‘collect’ and ‘runJob’ is greater than the ‘sortBy’, so the application is IO-bound.

b. Look up the hourly costs of the m6gd.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?

Ans:

The hourly costs of m6gd.xlarge instance is \$0.1808.

The elapsed time of completing reddit-5 is 4 minutes. (4 instances)

Hence the cost of processing this larger dataset is $\frac{4}{60} \times \$0.1808 \times 4 \times 10 \approx \$0.482$

In this case, we should use more executors as we can, in other words, there should be a large amount of repartitions. There may be several cores in one instance, so we should take a large repartition number, maybe a number greater than 160.