ANSWERS

Q1.

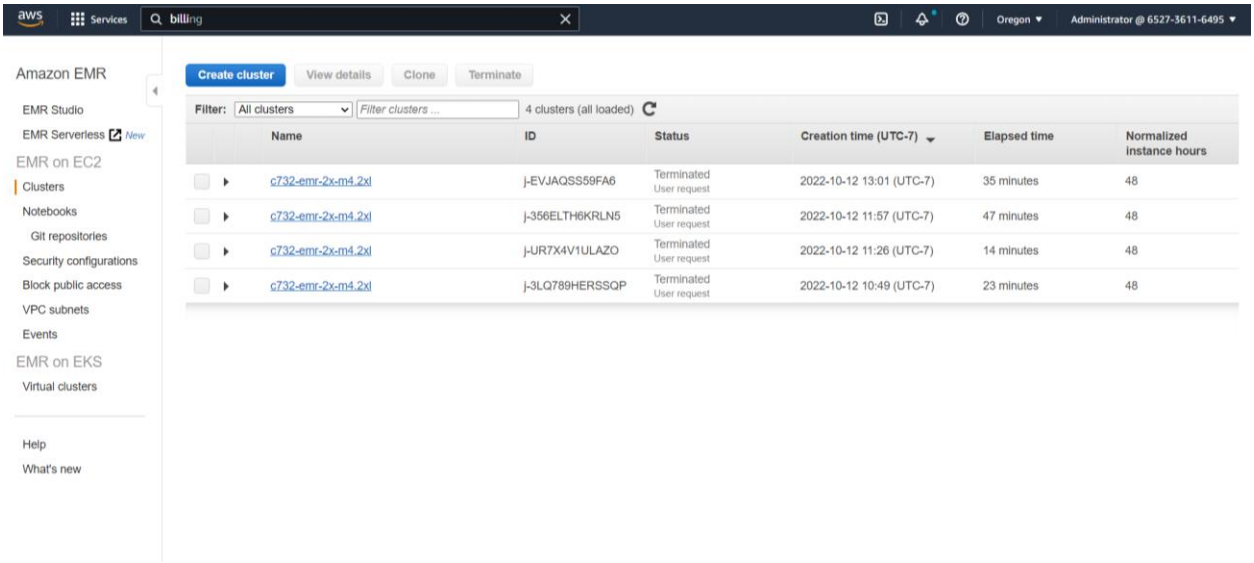

Figure1. Screenshot of all created cluster for assignment 5

Q2.

a) Input file size without prefilter at S3 is 2.6 MiB and with filter it was 97.7 KiB. So, S3 filtered (2662-97.7)/2662, that is, 96.33%
b) The filter, sorting (SQL where), file read options are done by both spark and S3.

Q3.

a)

The collect() job which included operations like reading text file and partitioning (Fig3) took the maximum time, making this application IO bound. The compute operations on the other hand took only 20-22s to complete. Even the last job, jobid5, took 1.3m which included writing the output to a file, again an IO task. (Fig4)

Spark 3.3.0-amzn-0 | Jobs | Stages | Storage | Environment | Executors | Reddit Average Spark application UI

Spark Jobs (?)

User: hadoop
Total Uptime: 6.2 min
Scheduling Mode: FIFO
Completed Jobs: 6

Event Timeline

Completed Jobs (6)

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
5	runJob at SparkHadoopWriter.scala:83 runJob at SparkHadoopWriter.scala:83	2022/10/12 20:26:33	1.3 min	2/2	32/32
4	sortBy at /mnt/tmp/spark-f1076cce-758b-48d8-998b-37cc803469c8/relative_score_bcast.py:48 sortBy at /mnt/tmp/spark-f1076cce-758b-48d8-998b-37cc803469c8/relative_score_bcast.py:48	2022/10/12 20:26:11	22 s	1/1	16/16
3	sortBy at /mnt/tmp/spark-f1076cce-758b-48d8-998b-37cc803469c8/relative_score_bcast.py:48 sortBy at /mnt/tmp/spark-f1076cce-758b-48d8-998b-37cc803469c8/relative_score_bcast.py:48	2022/10/12 20:25:50	21 s	1/1	16/16
2	collect at /mnt/tmp/spark-f1076cce-758b-48d8-998b-37cc803469c8/relative_score_bcast.py:44 collect at /mnt/tmp/spark-f1076cce-758b-48d8-998b-37cc803469c8/relative_score_bcast.py:44	2022/10/12 20:23:47	2.0 min	3/3	48/48
1	sortBy at /mnt/tmp/spark-f1076cce-758b-48d8-998b-37cc803469c8/relative_score_bcast.py:41 sortBy at /mnt/tmp/spark-f1076cce-758b-48d8-998b-37cc803469c8/relative_score_bcast.py:41	2022/10/12 20:23:27	20 s	1/1	16/16
0	sortBy at /mnt/tmp/spark-f1076cce-758b-48d8-998b-37cc803469c8/relative_score_bcast.py:41 sortBy at /mnt/tmp/spark-f1076cce-758b-48d8-998b-37cc803469c8/relative_score_bcast.py:41	2022/10/12 20:21:54	1.6 min	1/1	16/16

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

Figure2. List of all jobs with up time for relative score

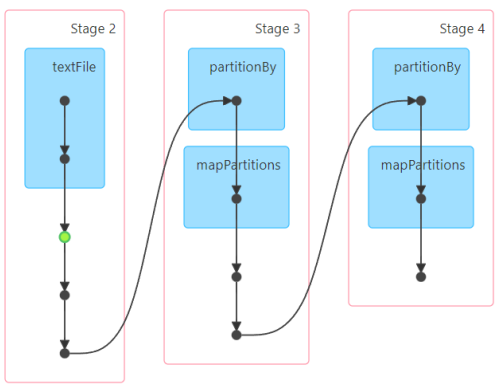

Figure3. Stages in job2

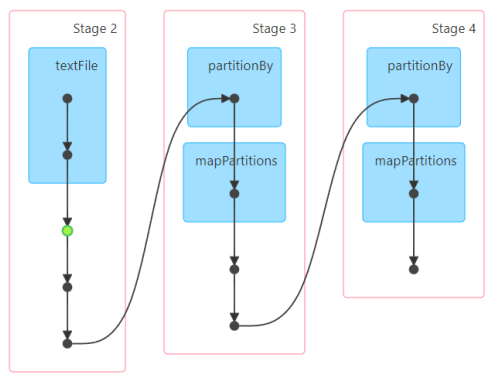

Figure4. Stages in job5

b)

m6gd.xlarge costs $0.1808 per hour, for 4 instances, (Memory = 16 GiB, Storage=1 x 237 NVMe SSD Network Storage=Up to 10 Gigabit). Therefore, the total costs would be 0.1808 * 4 * hours the cluster was running. If dataset reddit-5 was 10 times larger the costs would be 10 times more. This can be avoided if we can split the large reddit-5 data into smaller more manageable files to process on 16 instances.