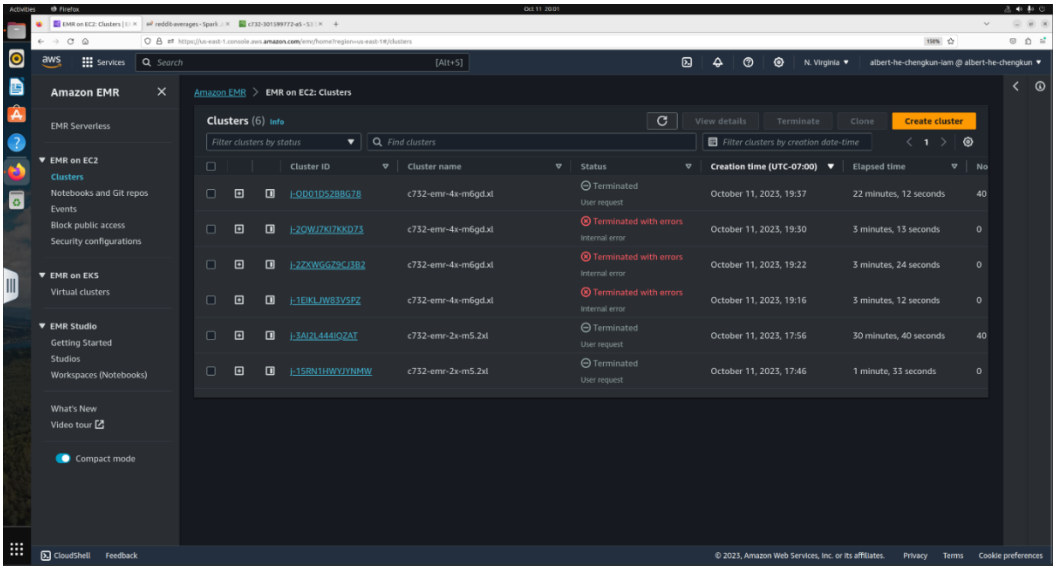
Name: He Chengkun
Student ID:

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.



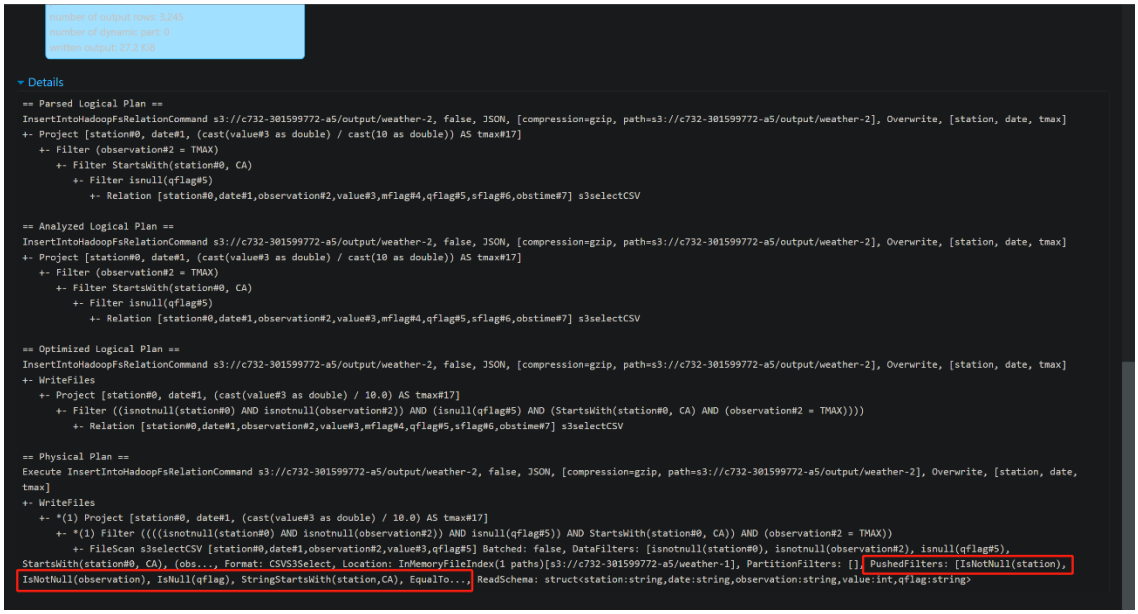
2. For Section 2:
   a. What fraction of the input file was prefiltered by S3 before it was sent to Spark?

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output
0	json at NativeMethodAccessorImpl.java:0 +details	2023/10/12 01:13:44	5 s	4/4	2.6 MiB	27.2 KiB

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output
0	json at NativeMethodAccessorImpl.java:0 +details	2023/10/12 01:23:31	4 s	4/4	97.7 KiB	27.2 KiB

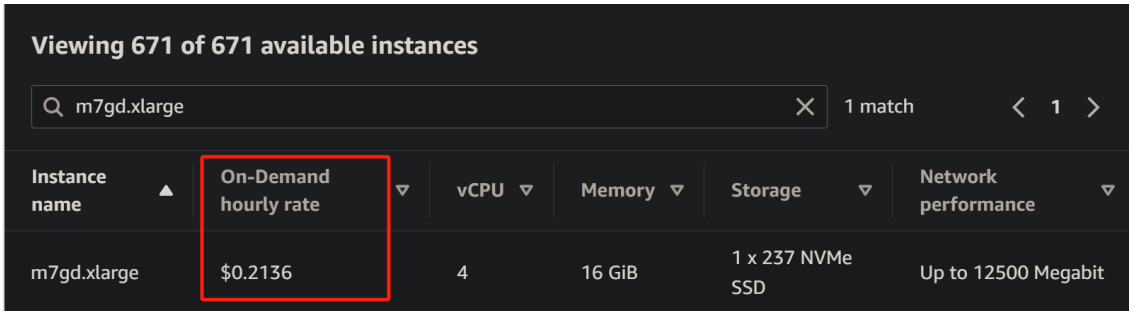
The original input is 2.4 MB, and the input prefiltered by S3 is 97.7 KB. Input size was reduced by ~2.30 MB.

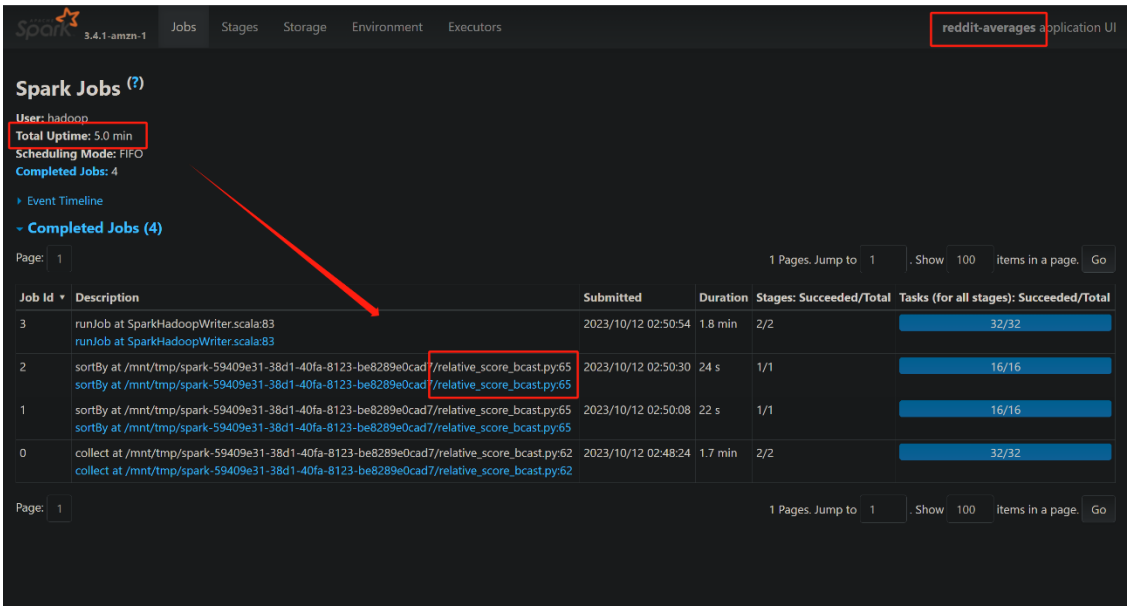
   b. Comparing the different input numbers for the regular version versus the prefiltered one, what operations were performed by S3 and which ones performed in Spark?



Take a look of the Spark History of weather ETL S3 select. From Spark History -> SQL/DataFrame -> Details - > Physical Plan -> PushedFilters, we can know that all the filtering operations were pushed down to S3.

3. 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?





Suppose that time cost grows linearly with the increment of data size, since processing reddit-5 costs 5 mins for spark on four m5gd.xlarge instances, it would reasonably cost less than 50 mins on four m7gd.xlarge instances for a dataset of 10-time size as reddit-5. Addition to the launching and terminating time, it would take about an hour from the start of cluster to its termination. Hence its hourly rate $0.2136 would be a reasonable estimation of the cost.

If we want to make full use of 16 instances, we can set instance count to 16 when creating / cloning the cluster.