**Assignment 5**

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

HY answers:

|  |  |
| --- | --- |
| Weather, no S3 filtering | Weather, with S3 filtering |
| A screenshot of a computer  Description automatically generated | A screenshot of a phone  Description automatically generated |

A screenshot of a computer code

Description automatically generated

2a. Based on the input size figures, 96.3% of the input file was prefiltered by S3.

(2.6 \* 1024 – 97.7) / 2.6 \* 1024 = 96.3%

2b. Seen from the different input numbers, as well as the physical plan with S3 filtering, most of the filtering processing were pushed down to S3, including IsNotNull(station), IsNotNull(observation), IsNull(qflag), and StringStartsWith(station,CA); while the filtering operation of selecting (observation#2 = TMAX) was performed in Spark.

1. For Section 3:

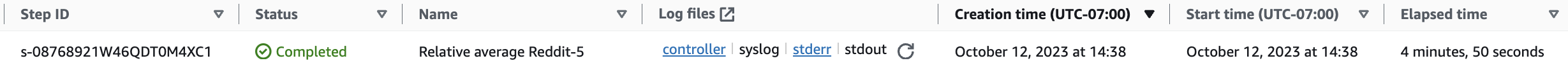
Look up the hourly costs of the m6gd.xlarge instance on the [EC2 On-Demand Pricing](https://aws.amazon.com/ec2/pricing/on-demand/) page. Estimate the cost of processing a dataset ten times as large as reddit-5 using just those 4 instances.

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

HY answers:

A screenshot of a computer

Description automatically generated



Cost estimation of processing a larger dataset: $0.583

* On-Demand hourly rate of one m6gd.xlarge instance: $0.1808
* The elapsed time of my step: 4 minutes, 50 seconds

(4 \* 60s + 50s) \* 10 / (60 \* 60s) \* $0.1808 \* 4 $0.583

When processing the larger dataset with 16 instances instead of 4, it is the best approach to make full use of the instances’ capacities by matching each paralleled thread a similar-sized task. Given that m6gd.xlarge can at most have 4 threads running at the same time, the larger dataset needs to be evenly distributed into 64 (4\*16) parts for the processing job.