1. The middle cluster was created incorrectly, which is why I have 3 clusters.

A screenshot of a computer

Description automatically generated

1. a)

Weather-1 input size: 2.6MiB

Weather-but-different input size: 327.6KiB

327.6KiB/2.6MiB = 123KiB

b)

The different file uses hive-partitioning. Hive-partitioning separates data into folders based on partition keys. When querying data, the system can skip over irrelevant partitions and only read the relevant ones, reducing the amount of data retrieved.

1. Total uptime for reddit-5 on cluster: 13min (3.9, 2.7, 2.8, 3.6mins per job)

Uptime on AWS for reddit-5: 2.8 min

Time to run a dataset 10x as large = 28min

M7gd.xlarge cost: 0.2136/hr

Cost = (28min/60min \* 4 instances) \*.2136/hr = ~ 40 cents