1. Observation: In most time, only 50% of CPU is used.

Analyse: There is still free CPU to be used. Therefore, we may increase ‘executor-cores’ and ‘spark.default.parallelism’ to increase the number of task threads and tasks.

1. Observation: There are three peaks in CPU utilization, reaching around 70% to 90%.

Analyse: The potential activities in the peaks are collectAsMap and mapPartitions.

1. Observation: The memory usage is stable at around 2300MB.

Analyse: Seems good.

1. Observation: The net utilisation is concentrated on several phases, but stable.

Analyse: Seems good.

1. Observation: The disk utilisation, I/O and block I/O process has five peaks.

Analyse: The peaks appear near the start of each circle (take -> count -> flatMap …). I will try Kryo serialisation on the cloud to check whether it helps.

1. Observation: The data per task is quite balanced, about 7000 to 9000 bytes per task. Therefore, there might not be the problem of data skew.
2. Probably useful settings

spark.shuffle.sort.bypassMergeThreshold