ICS 411: Introduction to Big Data Management

Assignment 3 Answer

**Question 3:** Comparison table and Observation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Counter Name | Case 1 | Case 2 | Case 3 | Case 4 |
| FILE: Number of bytes read | 404727 | 25821 | 404793 | 27093 |
| FILE: Number of bytes written | 1032227 | 274729 | 2259680 | 1506256 |
| HDFS: Number of bytes written | 20782 | 20782 | 21764 | 21764 |
| Total time spent by all map tasks (ms) | 5314 | 5442 | 5058 | 6457 |
| Total time spent by all reduce tasks (ms) | 5237 | 4458 | 342848 | 306540 |
| Combine input records | 0 | 20104 | 0 | 20104 |
| Combine output records | 0 | 1310 | 0 | 1370 |
| Reduce shuffle bytes | 20104 | 25821 | 404793 | 27093 |
| Reduce input records | 20104 | 1310 | 20104 | 1370 |
| Reduce output records | 1310 | 1310 | 1370 | 1370 |
| Spilled Records | 40208 | 2620 | 40208 | 2740 |
| Shuffled Maps | 1 | 1 | 12 | 12 |
| Merged Map outputs | 1 | 1 | 12 | 12 |
| Total committed heap usage (bytes) | 279384064 | 279384064 | 1238896640 | 1238896640 |

From the table data. We can observer that in the presence of combiner, the amount of bytes written to the file system, time spent by the reduce tasks, etc comes down. Combiner is like half reducer that performs on the mapper side and hence reduces the time/ data spent in shuffling. We can see that the Reduce Input records is drastically reduced from 20104 to 1310 in the presence of combiner and so the spilled records. Overall the combiner reduces the data shuffled and also time taken for the reducer.

When we observe the effect of having the partitioner, since there is 12 outputs, the number of bytes written in the filesystem, total spent time etc. everything increases, as in this case the outputs should be obtained for every specific month. Partitoner does not have much effect on the amount of shuffled bytes or reducer input records. Since there is 12 different reducers, partitioners help diving the data to each reducer. The increase in bytes written, memory all are due to the extra bytes that has to be handled for repeating the write for different months. When we compare having 12 reducers without using the partitioner and the one here using the partitioner, we can find that the partitioner usage betters the performance.