**Metropolitan State University**

**ICS 411-01 Big Data Storage and Processing**

**Fall 2019 - Assignment 3 Report**

**Case 1:**

Just Mapper and reducer.

**Case 2:**

Mapper, combiner, and reducer.

**Case 3:**

Mapper, partitioner, and reducer.

**Case 4:**

Mapper, combiner, partitioner, and reducer.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Counter Name** | **Case 1** | **Case 2** | **Case 3** | **Case 4** |
| FILE: Number of bytes read | 2268578 | 2261252 | 2268644 | 2261318 |
| FILE: Number of bytes written | 4823533 | 4809213 | 6400938 | 6388496 |
| HDFS: Number of bytes written | 2178867 | 2178867 | 2178867 | 2178867 |
| Total time spent by all map tasks (ms) | 2616 | 2469 | 2472 | 2315 |
| Total time spent by all reduce tasks (ms) | 2771 | 2507 | 64953 | 48472 |
| Combine input records | 0 | 20103 | 0 | 20103 |
| Combine output records | 0 | 20030 | 0 | 20030 |
| Reduce shuffle bytes | 2268578 | 2261252 | 2268644 | 2261318 |
| Reduce input records | 20103 | 20030 | 20103 | 20030 |
| Reduce output records | 20030 | 20030 | 20030 | 20030 |
| Spilled Records | 40206 | 40060 | 40206 | 40060 |
| Shuffled Maps | 1 | 1 | 12 | 12 |
| Merged Map outputs | 1 | 1 | 12 | 12 |
| Total committed heap usage (bytes) | 555220992 | 555745280 | 2337800192 | 2285895680 |

*(Explanations for the highlighted lines on the following page.)*

**Total time spent by all map tasks (ms)**

Cases 3 and 4 ran more quickly than the first two cases, mainly due to the partitioner. The fastest run used both the combiner and the partitioner. Just using the partitioner alone provided improvement by allocating a separate reducer for the assigned tasks coming out of the mappers. The combiner improves on this by organizing the data before it gets to the reducers by combining like keys and their corresponding values (counts).

**Total time spent by all reduce tasks (ms)**

Cases 3 and 4 took considerably longer due to the quanitity of reducers involved. The first two cases only used the default number, which is 1, whereas the last two cases used 12 (as defined in the driver class).

**Combine input records / Combine output records**

Only Case 2 and Case 4 have any combiner input records since those were the only cases for which a combiner was used.

**Reduce input records**

These rows were not highlighted in the table above, but worth noting that there was a small difference in the number of records run with and without a combiner. The combiner seems to have provided an improvement of 73 fewer records for the reducer to process.

**Shuffled Maps / Merged Map outputs**

These two counters reflect the number of reducers that ran during the job’s execution. The first two columns show a value of 1 representing the default number of reducers, and the last two columns show 12 – one reducer for each month of the year, with each reducer being assigned by the custom partitioner.