Aayush Shah
19BCE245
17 October 2022

Big Data Analytics

Practical 6

Aim

Analyse impact of different number of mapper and reducer on same definition.

Mappers

- Hadoop Mapper is a function or task that processes all input records from a file and produces output that may be used as input for Reducer.
- No. Of Mapper : Total data size/Input split size

1. By default, in Hadoop, the input size split is 128 MB. However, we can change the same. On reducing the size, the number of mappers will increase :

   set.mapreduce.input.fileinputformat.split.maxsize=100000;

   For example, for block size 100MB and input of 1GB, 10 mappers will be used.
2. We can also set the number of mappers as in the driver code *Job.setNumMapTask(5)* which sets 5 number of mappers.
3. While executing the job *hadoop jar /hirw-starterkit/mapreduce/stocks/MaxClosePrice-1.0.jar com.hirw.maxcloseprice.MaxClosePrice -D mapred.reduce.tasks=10 /user/hirw/input/stocks output/mapreduce/stocks*

- The right level of parallelism is 10-100 mappers per node; if the mappers are relatively small, then, maybe 300 mappers per node.
- If the number of mappers is too much or too less, it can slow the system down.

Reducers

- Reducer in Hadoop MapReduce reduces a set of intermediate values which share a key to a smaller set of values.

- In MapReduce job execution flow, Reducer takes a set of an intermediate key- value pair produced by the mapper as the input.
- As given on the website of Hadoop the ideal number of Reducers are *(0.95 or 1.75) * (nodes * number of mappers)*.

  1. We can also set the number of reducers as *Job.setNumReduceTask(5)* which set 5 number of reducers.
  2. Using command line *jar word_count.jar com.home.wc.WordCount /input /output -D mapred.reduce.tasks=20* which sets the number of the reducers to 20.

  Suppose there are 100 reduce slots available in the cluster.

- When we consider 0.95 as the factor, there will be 95 reducers and it means that no task will be waiting.
- This is to be done when all the tasks take equal number of times.

When we take 1.75 as the factor, 100 tasks will start simultaneously, while 75 will be in the queue.

- This will allow better load balancing as it would prevent bottleneck of the jobs.

Conclusion

We learned about the many ways we may adjust the amount of mappers and reducers and how this affects the job in general in this practical. When performing tasks with TBs of data, it is critical to select the appropriate amount of mappers and reducers.