18th October, 2018

Week-4 Report
D Swami

Project Title: On studying the performance of Hadoop MapReduce vs MPI for Aggregation Operations: A Big Data Challenge

The following project aimed at benchmarking various parameters of Map Reduce & MPI for parallel I/O. In the fourth week of the work, I have accomplished following tasks:

1) Debug Map Reduce to get rid of logical errors.
2) Soft test Map Reduce on 27 GB data.
3) Partial completion of MPI programs

Issues tackled in the current week:

1) Error: “Map Reduce failed due to NA”. Introduced checks for bad values into the Map Reduce Framework.
2) SkipBadRecord class is not available for the mapreduce api and hence a suggestion to others to use mapred framework instead.

Tasks for the upcoming week:

1) Compute wall clock time for 128 MB and 64 MB of block size in Map Reduce for the complete dataset.
2) Complete the remaining portion of the MPI program.
3) If time permits calculate wall clock time for varying Input-split size.

Expected Issues in the coming week:

1) Disk space related issues if the amount of disk space is not enough for Map Reduce to process 215 GB data. We have a current capacity of more than 380 GB. Hence, we don’t expect the same but also should not be ignored.
2) Time management: Processing 200 GB data would take a long time and requires uninterrupted processing.