16th October, 2018

Week-5 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 fifth week of the work, I have accomplished following tasks:

1) Computed wall clock time for 128 MB block, which is around 18 hours approximately using 32 MB split size and no compression. There were in all 6700+ tasks that composed of the complete job.
2) Completed 1st version of MPI program and its testing using Visual Studio with Intel parallel studio suite of compilers.
3) Testing basic programs related to I/O for MPI on Cedar clusters.
4) Concluded that YARN is a mess for interactive queries. It has too much of overhead associated with calling tasks.

Issues tackled in the current week:

1) Error: “Map Reduce failed due to NA”. Recent years had change the data dictionary and hence were not compatible for files from 2016 and 2017 especially.
2) Failed Map Reduce due to

Tasks for the upcoming week:

1) Compute wall clock time for 64 MB of block size in Map Reduce for the complete dataset.
2) Complete the 2nd version of the MPI program.
3) Calculate wall clock time for varying Input-split size (32 MB, 64 MB, 128 MB).
4) Calculate wall clock time for varying Compression format.

Expected Issues in the coming week:

1) Time management problems, a split size of 32 MB on 200 GB data will decompose the job into 6400+ tasks.
2) Time management: Processing 200 GB data would take a long time and requires uninterrupted processing.