16th October, 2018

Week-6 and Week 7 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 sixth and the seventh week of the work, I have accomplished following tasks:

1) Computed wall clock time for remaining Map Reduce Tests. Results in Graph below.
2) Tested MPI program with 5000 data points on SHARCNET.

Issues tackled in the current week:

1) Changed the usage of MPI_Scatter to MPI_BCast so as to save memory. This had to be done since if no of data points are not a multiple of no. of processors than it gets tricky.
2) SHARCNET has gcc 5.4 compiler which dosenot support the newer C++ routines like stof, stoi, and others. Solution use compile flags: “-lstdc++” and “-std=c++11”.
3) To review all the previous logs of execution of Map Reduce tasks shifted to Job-History sever which still uses YARN for tracking metrics.

Tasks for the upcoming week:

1) Upload data on Cedar cluster for MPI performance analysis.
2) Perform MPI on the uploaded data.
3) Relevant updates to MPI codes for further improvement.

Expected Issues in the coming week:

1) Debugging challenges and requesting right amount of resources from Cedar cluster.
2) Errands pertaining to data quality